IMPROVING THE EMPIRICAL RISK BOUND FOR PRIVATE LEARNING UNDER INTERPOLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Differentially Private Gradient Descent (DP-GD) provides a principled approach to training machine learning models while preserving data privacy by injecting Gaussian noise into model updates and clipping sample-wise gradients. Existing literature has established joint theoretical guarantees for differential privacy and excess empirical risk under classical assumptions such as (strong) convexity and the Polyak–Łojasiewicz (PL) condition, with excess empirical risk bounded by $\tilde{\mathcal{O}}(p/(n^2\epsilon^2))$ or worse. In this paper, we study DP-GD in the *interpolation regime*, a key setting in modern machine learning where the training data can be fit exactly. We show that interpolation enables substantially stronger optimization guarantees, transforming polynomial-type empirical-risk bounds into exponential-type bounds in p and n, while still preserving differential privacy. Specifically, leveraging a property of interpolation – *automatic gradients reduction*, we propose to synchronously decay the standard deviation of the injected Gaussian noise and the gradient clipping threshold, at an appropriate exponential rate. First, in the classical settings considered in prior work, we show that interpolation improves the empirical risk to $\mathcal{O}(\exp(-\tilde{\Theta}(n^2\epsilon^2/p)))$. We further extend our analysis to non-convex neural networks and establish the same exponential empirical risk bound, which is also preserved under stochastic training. Finally, empirical evaluations verify the effectiveness of the proposed hyperparameter decay mechanism and demonstrate improved convergence compared with standard DP-GD.

1 INTRODUCTION

In an era where vast amounts of private data, ranging from electronic health records and financial transactions to location traces from smartphones and online browsing histories, are aggregated to train deep neural networks and large-scale foundation models, protecting individual privacy has become both ethically critical and legally mandated. However, many machine learning models, including deep neural networks, are intrinsically vulnerable to privacy leakage under attacks (Carlini et al., 2021; Haim et al., 2022). To address this challenge, Differentially Private Gradient Descent (DP-GD) and its variants provide a standard mechanism to preserve differential privacy and protect individual data points during neural network training. Compared to standard gradient descent, these methods introduce two key modifications to the update step: injecting random Gaussian noise and clipping sample-wise gradients (Bassily et al., 2014; Abadi et al., 2016).

A rich literature has theoretically established that DP-GD can simultaneously provide data privacy and optimization guarantees. Specifically, under the classical settings of (strong) convexity or the Polyak–Łojasiewicz (PL) condition, with appropriate choices of algorithmic hyperparameters and training duration, the excess empirical risk can be bounded on the order of $\tilde{\mathcal{O}}(p/n^2)$ or $\tilde{\mathcal{O}}(\sqrt{p}/n)$ (see Table 1), where p is the number of model parameters and n is the number of training samples (Bassily et al., 2014; Wang et al., 2017; Yu et al., 2020; Fang et al., 2023). However, these analyses have several important limitations. First, they rely on classical assumptions that are generally not satisfied by neural networks. Second, the resulting convergence rates are slow, leading to pessimistic empirical risk bounds that decay only polynomially with p/n^2 and can therefore yield large final training errors that are insufficient for meaningful learning. In contrast, although plain gradient descent does not provide a privacy guarantee, it can converge at an exponential rate and correspondingly achieve an exponentially small empirical risk under both classical conditions, including strong convexity and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Summary of risk bounds of DP-GD analyses.

Analysis	Risk Bound	Conditions on Loss Function
Bassily et al. (2014)	$\tilde{\mathcal{O}}(\sqrt{p}/(n\epsilon))$	Convexity
Bassily et al. (2014)	$\tilde{\mathcal{O}}(p/(n^2\epsilon^2))$	Strong Convexity
Wang et al. (2017)	$\tilde{\mathcal{O}}(p/(n^2\epsilon^2))$	Polyak–Łojasiewicz (PL)
Yu et al. (2020)	$\tilde{\mathcal{O}}(p/(n^2\epsilon^2))$	Convexity + Expected Curvature
Fang et al. (2023)	$\tilde{\mathcal{O}}(\sqrt{p}/(n\epsilon))$	Convexity + Weak Growth
Song et al. (2021)	$\tilde{\mathcal{O}}(\sqrt{\mathrm{rank}}/(n\epsilon))^*$	Generalized Linear Model (Convexity)
Arora et al. (2022)	$\tilde{\mathcal{O}}(1/\sqrt{n} + 1/(n\epsilon)^{2/3})$	Generalized Linear Model (Convexity)
Bombari & Mondelli (2025)	$\tilde{\mathcal{O}}(d/(n\epsilon) + \sqrt{d/n} + \sqrt{n/d^{3/2}})^\dagger$	Random Feature Model (Convexity) + Interpolation
This work	$\mathcal{O}(\exp(-\tilde{\Theta}(n^2\epsilon^2/p)))$	(Generalized) Strong Convexity or PL + Interpolation
This work	$\mathcal{O}(\exp(-\tilde{\Theta}(n^2\epsilon^2/p)))$	Over-parameterized neural networks (PL* + Aiming + Interpolation)

the PL condition (Nesterov, 2013), as well as in neural networks (Allen-Zhu et al., 2019; Du et al., 2019a; Oymak & Soltanolkotabi, 2020; Liu et al., 2022; Zou et al., 2020).

In this paper, we study DP-GD in the *interpolation* regime and establish substantially stronger optimization guarantees, showing that DP-GD can achieve exponentially fast convergence and an improved empirical risk of exponential form. Interpolation refers to the setting in which the training data can be fit exactly, so that the minimum empirical loss is zero. This regime is particularly relevant to modern machine learning, where over-parameterized models with $p \geq n$ are common and can readily interpolate the training data. Perhaps surprisingly, despite achieving zero training error, interpolating models can exhibit strong generalization performance rather than the overfitting predicted by classical statistical theory (Zhang et al., 2017; Belkin et al., 2019). This phenomenon has motivated extensive theoretical studies of interpolation and its role in modern machine learning. In addition, interpolation is relevant beyond the over-parameterized regime: even when $p < n$, models may still achieve very small training loss, resulting in approximate interpolation.

Our approach leverages a fundamental property of interpolation, which we call *automatic gradients reduction*: as the training loss decreases, the magnitudes of the sample-wise gradients also tend to diminish. Motivated by this property and the exponential convergence of training loss in the interpolation regime, we propose to exponentially decay both the standard deviation of the injected Gaussian noise and the gradient clipping threshold, in tandem with the shrinking gradients. We refer to the resulting algorithm as *Noise-Decay DP-GD* (ND-DP-GD).

To show the improved empirical risk guarantee, we first consider the classical settings underlying existing theoretical analyses of DP-GD. We show that ND-DP-GD, with a constant learning rate η and exponentially decaying noise level and clipping threshold, converges exponentially fast and achieves an exponential empirical risk of order $\mathcal{O}(\exp(-\tilde{\Theta}(n^2\epsilon^2/p)))$ while maintaining differential privacy. This improves substantially upon the polynomial convergence and risk guarantees established in prior work. We then extend our analysis to neural networks, whose loss landscapes are non-convex and have been largely beyond the scope of existing DP-GD theory. We leverage recent advances in deep learning theory showing that, for sufficiently wide neural networks, the loss landscape can be characterized by the PL* condition (Liu et al., 2022), which is weaker than the standard PL condition, together with the Aiming condition (Liu et al., 2023). Under these conditions, we show that ND-DP-GD achieves the same exponential empirical risk guarantee with high probability over the random initialization of the network. See Table 1 for a comparison between our result and prior work. Finally, we show that this exponential optimization guarantee is preserved under stochastic training. Numerical evaluations further demonstrate that the proposed noise-decay mechanism outperforms standard DP-GD, consistent with our theoretical analysis.

We summarize our contributions as follows:

- We study differentially private optimization in the interpolation regime and introduce Noise-Decay DP-GD (ND-DP-GD), a variant of DP-GD that employs a synchronized decay mechanism for both the clipping threshold and the injected privacy noise.

$^*\mathrm{rank}$ refers to the rank of the feature matrix, which is upper-bounded by $\min(n, p)$.

$^\dagger d$ is the input data dimension, required to satisfy $d \ll n \ll d^{3/2}$ in (Bombari & Mondelli, 2025).

- Under interpolation and the classical convexity/PL-type settings, we prove that ND-DP-GD achieves an exponential empirical risk of order $\mathcal{O}(\exp(-\tilde{\Theta}(n^2\epsilon^2/p)))$ while preserving differential privacy. This substantially improves upon the existing risk bounds which are polynomial, e.g., $O(p/n^2)$.
- We further extend the analysis to the non-convex neural network settings, which is largely absent from prior DP-GD analyses. We show that the same exponential empirical risk of order $\mathcal{O}(\exp(-\tilde{\Theta}(n^2\epsilon^2/p)))$ still holds. We also show that similar results extend to stochastic training.

Notation. Throughout this paper, for a positive integer n, we use $[n]$ to denote the set $\{1, 2, \ldots, n\}$. The standard Euclidean norm (L_2-norm) of a vector is denoted by $\|\cdot\|$. For a differentiable function f, ∇f denotes its gradient. We use standard asymptotic notations $\mathcal{O}(\cdot)$ and $\Theta(\cdot)$ to absorb constant factors, while $\tilde{\mathcal{O}}(\cdot)$ and $\tilde{\Theta}(\cdot)$ are used to further hide poly-logarithmic factors. Additionally, $\mathbb{E}[\cdot]$ and $\mathbb{P}(\cdot)$ denote the expectation and probability with respect to the underlying randomness.

Organization. The paper is organized as follows: In Section 2, we review DP-GD and its theoretical guarantees in classical settings, as well as the concept of interpolation. In Section 3, we introduce the ND-DP-GD and establish its exponential empirical risk bound in the classical settings. In Section 4, we extend the theory to non-convex over-parameterized neural networks. Section 5 demonstrates that this theoretical guarantee is preserved under stochastic training. Appendix I provides numerical evaluations to demonstrate the consistency of our theory.

1.1 Related Work

Excess Empirical Risk Bounds. Excess empirical risk in differentially private optimization has been extensively studied for DP-GD and its stochastic variants (DP-SGD) (Song et al., 2013; Dwork & Roth, 2014; Abadi et al., 2016). For convex and strongly convex losses, classical results establish bounds with polynomial dependence on the parameter dimension p and sample size n, including $\tilde{\mathcal{O}}(p/(n^2\epsilon^2))$ and $\tilde{\mathcal{O}}(\sqrt{p}/(n\epsilon))$ (Bassily et al., 2014; Yu et al., 2020; Feldman et al., 2020; Fang et al., 2023). These guarantees have been extended to the global PL condition (Wang et al., 2017), while retaining polynomial dependence. For parameter-linear models, such as generalized linear models (GLMs) and random feature models, subsequent work reduces the dimension dependence (Song et al., 2021; Arora et al., 2022; Bombari & Mondelli, 2025), but the resulting bounds remain polynomial. Moreover, these analyses do not directly apply to the nonconvex optimization dynamics of over-parameterized neural networks.

Dynamic Clipping and Noise Scheduling. A core operation in DP-GD is gradient clipping. Standard static clipping distorts the optimization trajectory and introduces a non-vanishing bias term (Chen et al., 2020; Fang et al., 2023). Dynamic strategies designed to mitigate this issue generally fall into two primary categories. The first category relies on data-driven adaptive clipping (Pichapati et al., 2019; Andrew et al., 2021; Bu et al., 2023), which adjusts the threshold based on current gradient statistics. Recent advancements further explore gradient norm distribution estimation (e.g., DC-SGD (Wei et al., 2025)) and per-sample adaptive weights (e.g., DP-PSAC (Xia et al., 2023)). The second category explores heuristic dynamic decay schedules across training iterations. Methods in this category attempt to progressively decay noise scales, privacy budgets, or clipping thresholds based on training stages, such as empirical gradient decay curvatures in federated settings (Mao et al., 2023), adapted privacy budgets (Yu et al., 2019; Deng et al., 2022), or manual noise decay (Du et al., 2021; Zhang et al., 2021). While both categories demonstrate improved empirical performance, their theoretical guarantees only bound the gradient norm to establish convergence to an approximate stationary point, without providing guarantees on the empirical risk. In contrast, our analysis establishes that under interpolation, a synchronized decay mechanism guarantees an improved empirical risk with exponential dependence on p and n.

2 Background

Given a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ of size n, with inputs $\mathbf{x}_i \in \mathbb{R}^d$ and labels $y_i \in \mathbb{R}$, and a model f parameterized by $\mathbf{w}$, we consider the Empirical Risk Minimization (ERM) problem:

$$\text{minimize } \mathcal{L}(\mathbf{w}; \mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(\mathbf{w}; \mathbf{x}_i), y_i). \tag{1}$$

We assume that the individual loss $\ell(f(\mathbf{w};\mathbf{x}_i),y_i)$ is non-negative and denote it by $\ell_i(\mathbf{w})$ for short. Most commonly used loss functions, such as the square loss, $\ell(f(\mathbf{w};\mathbf{x}_i),y_i) \triangleq (f(\mathbf{w};\mathbf{x}_i)-y_i)^2$, and the cross-entropy loss, satisfy this non-negative assumption.

When the training dataset $\mathcal{D}$ contains sensitive information, **differential privacy** (Dwork et al., 2006) quantifies the privacy leakage of the learning algorithm. This framework is formally defined over adjacent datasets, where two datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ are considered adjacent if they differ by one individual's record.

Definition 2.1 ((ϵ,δ)-Differential Privacy)**.** Given $\epsilon>0$ and $\delta\in[0,1)$, a randomized mechanism $\mathcal{M}:\mathcal{X}^n\to\mathcal{R}$ satisfies (ϵ,δ)-differential privacy if for any two adjacent datasets $\mathcal{D}_1,\mathcal{D}_2\in\mathcal{X}^n$ and any subset of outputs $S\subseteq\mathcal{R}$ it holds that:

$$\mathbb{P}[\mathcal{M}(\mathcal{D}_1)\in S]\le e^{\epsilon}\mathbb{P}[\mathcal{M}(\mathcal{D}_2)\in S]+\delta. \tag{2}$$

The **Differentially Private Gradient Descent (DP-GD)** algorithm trains models aiming for protecting sensitive data via differential privacy guarantees. It differs from plain gradient descent via two operations: *noise injection* and *gradient clipping*. At each iteration t, DP-GD clips each sample-wise gradient $\nabla\ell_i(\mathbf{w}_t)\triangleq\nabla\ell(\mathbf{x}_i,y_i;\mathbf{w}_t), i\in[n]$ so that its norm does not exceed a clipping threshold C:

$$\tilde{g}_i^{(t)}=\nabla\ell_i(\mathbf{w}_t)/\max\left(1,\frac{\|\nabla\ell_i(\mathbf{w}_t)\|}{C}\right). \tag{3}$$

The model parameters are updated using the average of these clipped sample-wise gradients and an injected noise randomly sampled from isotropic Gaussian:

$$\mathbf{w}_{t+1}=\mathbf{w}_t-\frac{\eta}{n}\left(\sum_{i=1}^{n}\tilde{g}_i^{(t)}+\zeta^{(t)}\right),\quad \zeta^{(t)}\sim\mathcal{N}(0,C^2\sigma^2\boldsymbol{I}_p), \tag{4}$$

where η is the learning rate, $\boldsymbol{I}_p$ is the p-dimensional identity matrix, and σ is a noise scaling factor that controls the magnitude of the injected noise.

In general, the hyperparameters, clipping threshold C and noise scale $C\sigma$, can be time-dependent and pre-scheduled. Variants of DP-GD with specific schedules for these hyperparameters have been proposed (Andrew et al., 2021; Du et al., 2021). However, to the best of our knowledge, theoretical guarantees on the empirical risk for these variants have not yet been established.

Existing theoretical guarantees of DP-GD are mainly obtained under the classical settings of convex loss functions. In such settings, a substantial body of theoretical work establishes guarantees on both differential privacy and excess risk for DP-GD with constant hyperparameters σ and C (Bassily et al., 2014; Yu et al., 2020; Fang et al., 2023) or in linear-type models (Song et al., 2021; Arora et al., 2022; Bombari & Mondelli, 2025). However, the injected noise perturbs the optimization trajectory, resulting in a privacy-utility tradeoff (Chaudhuri et al., 2011; Song et al., 2021; Bombari & Mondelli, 2025). This tradeoff is characterized by the following bounds.

Theorem 2.2 (Privacy Guarantee, from Abadi et al. (2016))**.** *There exist constants c_1 and c_2 such that given T training steps, for any $\epsilon<c_1T$, the DP-GD algorithm satisfies (ϵ,δ)-differential privacy for any $\delta>0$, if $\sigma\ge c_2\frac{\sqrt{T\ln(1/\delta)}}{\epsilon}$.*

Typically, the privacy budget ϵ is treated as a constant (e.g., $\epsilon=1$) and δ is chosen to be $o(1/n)$.

Theorem 2.3 (Empirical Risk Guarantee, from Bassily et al. (2014) & Yu et al. (2020))**.** *Suppose $\ell_i(\mathbf{w})$ is smooth, strongly convex, and Lipschitz continuous over a bounded convex domain $\mathcal{W}\subset\mathbb{R}^p$ for all $i\in[n]$. Under the same noise setting as in Theorem 2.2 and with appropriate choices of the learning rate η and the number of iterations T, the excess empirical risk satisfies the following bound:*

$$\mathbb{E}[\mathcal{L}(\mathbf{w}_T)-\mathcal{L}(\mathbf{w}^*)]=\tilde{\mathcal{O}}\left(\frac{p\ln(1/\delta)}{n^2\epsilon^2}\right). \tag{5}$$

Under global strong convexity or global PL condition, the excess empirical risk is generally bounded by $\tilde{\mathcal{O}}(p/n^2)$. In simpler settings of linear models (e.g., generalized linear models, random feature models), this bound can be independent of the dimension p, achieving rates of $\tilde{\mathcal{O}}(1/n)$ or $\tilde{\mathcal{O}}(1/\sqrt{n})$, as summarized in Table 1.

2.1 Over-parameterization and Interpolation

Over-parameterization and interpolation have emerged as central phenomena in modern machine learning. As machine learning models, particularly neural networks, continue to grow in size, the number of model parameters p often exceeds the number of training samples n, resulting in an over-parameterized optimization problem. For example, ResNet-50, which has approximately 25 million parameters, can be trained on CIFAR-10, with $n = 50{,}000$ training samples, or ImageNet, with approximately $n = 1.28$ million training samples. In over-parameterized settings, a model can often *interpolate* the training data, as defined below.

Definition 2.4 (Interpolation)**.** A model interpolates the training data if there exists $\mathbf{w}^* \in \mathbb{R}^p$ such that $f(\mathbf{w}^*; \mathbf{x}_i) = y_i, \forall i \in [n]$.

When interpolation occurs, each individual loss is zero, i.e., $\ell_i(\mathbf{w}^*) = 0$, and consequently the total training loss also achieves its global minimum: $\mathcal{L}(\mathbf{w}^*) = 0$.

Interpolation challenges the classical view of overfitting. In classical statistical learning, fitting the training data too closely is generally associated with overfitting and poor generalization; in contrast, it has been widely observed that machine learning models can interpolate the training data while simultaneously achieving good generalization performance (Zhang et al., 2017; Belkin et al., 2019). This observation has motivated extensive theoretical studies of optimization and generalization in the over-parameterized regime.

Notably, interpolation is not unique to neural networks and also arises in classical machine learning models. For example, kernel methods without ridge regularization can interpolate the training data, as can random feature models when the number of random features exceeds the number of training samples. More broadly, interpolation approximately holds whenever the total training loss is sufficiently small. Note that, for a machine learning model to meaningfully learn from the training data, achieving a small training loss is often a necessary condition.

In the interpolation regime and using a random neural feature model which is linear in parameters, a recent work by Bombari & Mondelli (2025) has derived a polynomial risk bound of $\tilde{\mathcal{O}}(d/(n\epsilon) + \sqrt{d/n} + \sqrt{n/d^{3/2}})$, where the data dimension d is required to be $d \ll n \ll d^{3/2}$.

In the rest of the paper, we will leverage properties of interpolation to derive an improved empirical risk bound exponential in p and n, in both the classical settings of strongly convex and global PL and the modern non-convex setting for neural networks.

3 Exponential empirical risk in classical settings

In this section, we focus on the classical settings of strong convexity and the global PL condition, which underlie prior analyses of DP-GD. We establish an exponential empirical risk bound for DP-GD in the interpolation setting by employing a synchronized exponential decay schedule for the standard deviation of injected noise $C\sigma$ and clipping threshold C.

3.1 Automatic gradients reduction and Noise-decay DP-GD

We begin with a property of the sample-wise gradients under interpolation. First, we need the smoothness of the individual loss functions $\ell_i(\mathbf{w})$, which is a common assumption in the literature and is also assumed in prior DP-GD analyses (Bassily et al., 2014; Wang et al., 2017; Fang et al., 2023).

Assumption 3.1 (Smoothness)**.** Each individual loss function $\ell_i(\mathbf{w})$ is continuously differentiable and β-smooth. That is, $\|\nabla \ell_i(\mathbf{w}_1) - \nabla \ell_i(\mathbf{w}_2)\| \leq \beta \|\mathbf{w}_1 - \mathbf{w}_2\|$ for all $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W} \subset \mathbb{R}^p$ and all $i \in [n]$, where $\mathcal{W}$ is a domain of interest in the parameter space.

Under the smoothness of the individual loss functions, interpolation directly implies that the individual gradient norms are controlled by the empirical loss. Consequently, as the empirical loss decreases toward zero during training, the individual gradients automatically shrink toward zero; in particular, $\|\nabla \ell_i(\mathbf{w})\| \to 0$ as $\mathbf{w} \to \mathbf{w}^*$. We refer to this phenomenon as *automatic gradients reduction*.

Lemma 3.2 (Automatic Gradients Reduction)**.** *Suppose the optimization problem, Eq. (1), is in the interpolation regime and Assumption 3.1 holds. Then, the expected squared norm of the sample-wise*

Algorithm 1 Noise-Decay DP-GD (ND-DP-GD)

1: **Input:** Number of iterations T, initial clipping threshold C_0, decay rate q, learning rate η, noise scaling factor σ, initialization $\mathbf{w}_0$.
2: **for** $t \leftarrow 0, \ldots, T-1$ **do**
3: Set the clipping threshold $C_t = C_0 q^t$
4: For each $i \in [n]$, compute sample-wise gradient $g_i^{(t)} = \nabla \ell_i(\mathbf{w}_t)$
5: Clip gradients using C_t: $\tilde{g}_i^{(t)} = g_i^{(t)} / \max(1, \frac{\|g_i^{(t)}\|}{C_t})$
6: Update: $\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{\eta}{n}\left(\sum_{i=1}^n \tilde{g}_i^{(t)} + \zeta^{(t)}\right)$, where $\zeta^{(t)} \sim \mathcal{N}(0, C_t^2\sigma^2 \boldsymbol{I}_p)$
7: **end for**
8: **Return:** $\mathbf{w}_T$

gradients is bounded by the empirical loss:

$$\mathbb{E}_{i\in[n]}[\|\nabla \ell_i(\mathbf{w})\|^2] \leq 2\beta \mathcal{L}(\mathbf{w}). \tag{6}$$

Remark 3.3. The automatic gradients reduction above is also a generalization of the *automatic variance reduction* property of stochastic gradients (Liu & Belkin, 2020), which has been established only for quadratic loss under interpolation. The proof of the lemma is adapted from classical gradient bounds under convex optimization (Bubeck, 2015), and is related to the weak growth condition in the literature (Vaswani et al., 2019; Fang et al., 2023). We include a self-contained proof for the preceding lemma in Appendix A.

This gradient reduction property reveals a fundamental limitation of standard DP-GD, in which the noise level $C\sigma$ and clipping threshold C are typically kept fixed throughout training, as in most existing analyses of DP-GD (Bassily et al., 2014; Wang et al., 2017; Fang et al., 2023). Recall that Gaussian noise is injected to mask the gradient associated with each individual sample, which may contain sensitive information, while gradient clipping limits the influence of any individual training example. As training progresses, the empirical loss is expected to decrease, leading to a corresponding reduction in sample-wise gradients due to automatic gradients reduction. Consequently, those fixed hyperparameters can become increasingly mismatched with the scale of the gradients at later stages of training. In particular, the fixed Gaussian noise can eventually become disproportionately large relative to the shrinking gradients, introducing excessive perturbation and slowing convergence, which can lead to pessimistic empirical-risk guarantees. Meanwhile, as the sample-wise gradients shrink, a fixed clipping threshold becomes increasingly loose relative to their scale, rendering the large threshold redundant. This mismatch suggests that the noise scale and clipping threshold should adapt to the evolving magnitude of the gradients throughout training.

To address this issue, we propose synchronously decaying the noise scale and the clipping threshold, in tandem with the decay of sample-wise gradients during training. It is well known that, under convexity and interpolation, the training loss of plain GD converges exponentially fast. This motivates us to adopt an exponential decay schedule. Specifically, we introduce a decay rate $0 < q < 1$ and set the clipping threshold to $C_t = q^t C_0$ and the noise variance to $C_t^2\sigma^2$. See Algorithm 1 for details.

Remark 3.4. Exponential decay schedules for these two hyperparameters have been empirically explored in the literature (Du et al., 2021), but not in a synchronized manner. Moreover, existing analysis only establishes algorithmic stabilization, without providing empirical risk guarantees.

3.2 Exponential Empirical Risk Bound

We now show that, under an appropriate choice of hyperparameters, the empirical risk can be improved to an exponential form in terms of p and n while maintaining differential privacy, when interpolation holds under either (generalized) strong convexity or the PL condition.

Assumption 3.5 (Strong Convexity)**.** The empirical loss $\mathcal{L}$ is μ-strongly convex. Namely, for all $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W} \subset \mathbb{R}^p$, the following hold:

$$\mathcal{L}(\mathbf{w}_1) \geq \mathcal{L}(\mathbf{w}_2) + \langle \nabla \mathcal{L}(\mathbf{w}_2), \mathbf{w}_1 - \mathbf{w}_2 \rangle + \frac{\mu}{2}\|\mathbf{w}_1 - \mathbf{w}_2\|^2. \tag{7}$$

The following theorem states that, under strong convexity and interpolation, decaying the clipping threshold and noise variance at a synchronized exponential rate q can achieve a tighter empirical risk.

Theorem 3.6 (Empirical Risk of ND-DP-GD under Strong Convexity)**.** *Suppose the optimization problem in Eq. (1) is in the interpolation regime, and Assumptions 3.1 and 3.5 hold. For any $\delta_c \in (0,1)$ and given privacy parameters ϵ, δ, define $K = \sigma\left(\sqrt{p} + \sqrt{2\ln\left(\frac{2T}{\delta_c}\right)}\right)$. Set the hyperparameters as $\sigma = \frac{\sqrt{32T\ln(1/\delta)}}{\epsilon}$, $C_0 = \beta\sqrt{\frac{2}{\mu}\mathcal{L}(\mathbf{w}_0)}$, $q = \sqrt{1-\frac{\mu\eta}{2}}$, and $\eta = \min\left\{\frac{1}{\beta}, \frac{\mu^2 n^2}{4\beta^3 K^2}\right\}$. When the algorithm runs $T = \tilde{\Theta}(n^2\epsilon^2)$ steps, with probability at least $1-\delta_c$, the empirical risk of ND-DP-GD is bounded by:*

$$\mathcal{L}(\mathbf{w}_T) \leq \exp\left(-\tilde{\Theta}\left(\frac{n^2\epsilon^2}{p}\right)\right)\mathcal{L}(\mathbf{w}_0). \tag{8}$$

Proof. See proof in Appendix C. □

The empirical risk bound established in Theorem 3.6 is exponential in $\frac{n^2}{p}$, which is a substantial improvement over the bounds of prior analyses which are often polynomial, $\tilde{\mathcal{O}}(\frac{p}{n^2})$ or worse. In addition, the optimization guarantee established in the above theorem holds with high probability over the randomness of the algorithm, rather than only in expectation.

In principle, strong convexity is compatible with interpolation. However, strong convexity generally does not hold in over-parameterized settings, where the set of global minimizers can be high-dimensional and forms an affine subspace of dimension $p-n$ under convexity. To accommodate over-parameterization, we consider a *generalized strong convexity*: strong convexity transverse to the solution set and flatness parallel to the solution set. More formally, let $\mathcal{S}$ denote the affine subspace of global minimizers, and let $\Pi_{\mathcal{S}^\perp}$ be the orthogonal projection onto its orthogonal complement, the generalized strong convexity requires $\mathcal{L}(\mathbf{w}_1) \geq \mathcal{L}(\mathbf{w}_2) + \langle\nabla\mathcal{L}(\mathbf{w}_2), \Pi_{\mathcal{S}^\perp}(\mathbf{w}_1-\mathbf{w}_2)\rangle + \frac{\mu}{2}\|\Pi_{\mathcal{S}^\perp}(\mathbf{w}_1-\mathbf{w}_2)\|^2$. Over-parameterized (generalized) linear models and random feature models provide examples that fall into this category. It is easy to see that the loss function is constant along directions parallel to $\mathcal{S}$; hence, gradients are perpendicular to $\mathcal{S}$ and weight updates parallel to $\mathcal{S}$ do not affect the loss $\mathcal{L}$ and have no impact on the optimization trajectory. Therefore, Theorem 3.6 can be extended straightforwardly to the generalized strong convexity; we omit the proof. The result of Theorem 3.6 also extends to the Polyak–Łojasiewicz (PL) condition. We detail this extension in Appendix C.

The following theorem states that in the same algorithmic setting, the privacy is also guaranteed.

Theorem 3.7 (Privacy Guarantee)**.** *For any $\delta \in (0,1)$ and $\epsilon \in (0, 8\ln(1/\delta))$, if the total iterations T and noise scaling factor σ satisfy $\sigma \geq \frac{\sqrt{32T\ln(1/\delta)}}{\epsilon}$, then Algorithm 1 is (ϵ,δ)-differentially private.*

Proof. See proof in Appendix D. □

Remark 3.8. An alternative formulation using Gaussian Differential Privacy (GDP) (Dong et al., 2022) provides a different privacy accounting mechanism. However, as this analysis does not alter the asymptotic empirical risk bound, we retain the classic Moments Accountant in Theorem 3.7 to maintain consistency with the literature. The detailed GDP analysis is provided in Appendix E, which also serves as a foundation for the stochastic case in Section 5.

4 Empirical Risk for Over-parameterized Neural Networks

In this section, we extend the exponential empirical risk bound to the non-convex over-parameterized neural networks. Specifically, we consider the square loss as the optimization objective function. As discussed in Section 2, over-parameterization permits neural networks to interpolate the training data (Zhang et al., 2017). Under this setting, the set of global minimizers $\mathcal{S} = \{\mathbf{w} \in \mathbb{R}^p : \mathcal{L}(\mathbf{w}) = 0\}$ forms a lower-dimensional continuous manifold rather than a single point (Liu et al., 2022). For any parameter vector $\mathbf{w}$, we define $\bar{\mathbf{w}} = \mathrm{proj}_{\mathcal{S}}(\mathbf{w})$ as its projection onto $\mathcal{S}$, and the distance to the solution manifold as $\mathrm{dist}(\mathbf{w}, \mathcal{S}) = \|\mathbf{w} - \bar{\mathbf{w}}\|$.

The non-convexity of neural network loss landscapes introduces extra difficulties for optimization analysis. In DP-free scenarios, based on analysis of neural tangent kernel (NTK) (Jacot et al., 2018), the literature has established that plain gradient descent and SGD can converge exponentially fast

to global minima. Key to these analyses are the approximately linear training dynamics (Lee et al., 2019) and the following geometric properties of the neural network loss landscape:

Theorem 4.1 (Liu et al. (2022; 2023))**.** *Consider a neural network of depth L with randomly initialized weights $\mathbf{w}_0$. Let λ_0 denote the bottom eigenvalue of NTK. Suppose $\lambda_0 > 0$. If the width m of the network satisfies $m = \tilde{\Omega}(nR^{6L+2}/(\lambda_0 - \mu)^2)$ with $R = \Omega(1/\sqrt{\lambda_0})$, then, for the square loss function $\mathcal{L}$, there exists a global minimizer $\mathbf{w}^*$ within the ball $\mathcal{B}_R(\mathbf{w}_0)$ and the following conditions hold over $\mathcal{B}_R(\mathbf{w}_0)$: there exist constants $\mu > 0$ and $\theta > 0$, such that*

$$\mathbf{PL}^*: \ \|\nabla\mathcal{L}(\mathbf{w})\|^2 \geq \mu\mathcal{L}(\mathbf{w}), \ \ \mathbf{Aiming}: \langle\nabla\mathcal{L}(\mathbf{w}), \mathbf{w} - \bar{\mathbf{w}}\rangle \geq \theta\mathcal{L}(\mathbf{w}), \ \ \forall\mathbf{w} \in \mathcal{B}_R(\mathbf{w}_0). \tag{9}$$

Remark 4.2. It has been shown that λ_0 is positive for ReLU neural networks, if the input data does not degenerate, for example see (Du et al., 2019b).

Remark 4.3. Although the above theorem requires sufficiently large width for the PL* and Aiming conditions to hold, empirical evidence suggests that these conditions often hold along the optimization trajectory even at practical widths. Specifically, Liu et al. (2023) provides numerical evidence supporting the Aiming condition. The PL* condition has also been frequently observed in practical over-parameterized training, as evidenced by the non-vanishing gradient norm and continued decrease in the loss when the loss remains non-zero.

While the PL* condition controls the magnitude of the gradient, preventing the training algorithm from stalling before reaching a global minimum, the Aiming condition controls its direction by requiring the negative full-batch gradient to point toward the solution set. Together, these conditions ensure that gradient descent makes progress both in magnitude and in the appropriate direction. A key distinction between the PL* condition and the standard PL condition is that PL* only requires the PL inequality to hold locally, e.g., within a ball $\mathcal{B}_R(\mathbf{w}_0)$, whereas the standard PL condition requires it to hold globally. For neural networks, the PL inequality generally does not hold over the entire parameter space. However, it can hold locally within a sufficiently large ball, which has been shown to contain the optimization trajectory of gradient descent. The Aiming condition, although it is often not stated explicitly in convex optimization, can be inferred from convexity, and plays an important role in convergence analysis under convexity, ensuring that the negative gradient provides a descent direction. However, the locality of these two properties of the neural network loss landscape introduces challenges in showing convergence of ND-DP-GD on neural networks. Leveraging the exponential decay of the injected noise scale, we are able to prove that the optimization trajectory remains in the local region with high probability, and have the following theorem:

Theorem 4.4 (Empirical Risk of ND-DP-GD on NNs)**.** *Suppose the optimization problem in Eq. (1) is in the interpolation regime, Assumption 3.1 holds, and μ-PL* and θ-Aiming conditions hold within a ball $\mathcal{B}_R(\mathbf{w}_0)$. For any $\delta_c \in (0,1)$ and given privacy budget $\epsilon > 0, \delta \in [0,1)$, let $\alpha = \frac{\mu}{4}$, $\gamma = \eta(\theta - \beta\eta)\alpha$ and $K = \sigma\left(\sqrt{p} + \sqrt{2\ln\left(\frac{2T}{\delta_c}\right)}\right)$. Set the hyperparameters of ND-DP-GD as $\sigma = \frac{\sqrt{32T\ln(1/\delta)}}{\epsilon}$, $C_0 = \beta\sqrt{\frac{2\mathcal{L}(\mathbf{w}_0)}{\alpha}}$, $q = \sqrt{1 - \frac{\gamma}{2}}$, and $\eta = \min\left\{\frac{\theta}{2\beta}, \frac{\theta\alpha n^2}{8\beta^2K^2}\right\}$. If the ball radius $R \geq \frac{2\eta C_0}{1-q}\left(1 + \frac{K}{n}\right)$, when the algorithm runs $T = \tilde{\Theta}(n^2\epsilon^2)$ steps, with probability at least $1 - \delta_c$, the empirical risk is bounded by:*

$$\mathcal{L}(\mathbf{w}_T) \leq \frac{\beta}{\alpha}\exp\left(-\tilde{\Theta}\left(\frac{n^2\epsilon^2}{p}\right)\right)\mathcal{L}(\mathbf{w}_0). \tag{10}$$

Note that the differential privacy guarantee established in Theorem 3.7 is independent of the loss landscape and applies directly to this setting.

In Appendix I, we empirically evaluate ND-DP-GD and demonstrate its improved convergence compared with standard DP-GD.

5 Extension to Stochastic Training

In practice, deep neural networks are usually trained using stochastic gradients rather than full-batch. Our exponential empirical risk bound extends to the stochastic setting via Poisson subsampling (Chua et al., 2024): ND-DP-SGD. Specifically, ND-DP-SGD follows Algorithm 1 except that the

parameters are updated using a mini-batch $\mathcal{B}_t$ of data points:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{\eta}{B_s}\left(\sum_{i\in\mathcal{B}_t}\tilde{g}_i^{(t)} + C_t\sigma\zeta_s^{(t)}\right). \tag{11}$$

At each iteration t, each data point is selected independently with probability q_s to form a mini-batch index set $\mathcal{B}_t \subseteq [n]$, yielding an expected batch size of $B_s = nq_s$.

We establish the privacy guarantee for ND-DP-SGD based on the property of privacy amplification by subsampling. Specifically, we extend the Gaussian Differential Privacy (GDP) framework detailed in Appendix E to incorporate the Poisson subsampling mechanism.

Theorem 5.1 (Privacy Guarantee for ND-DP-SGD)**.** *For any $\epsilon > 0$ and $\delta \in (0, 0.5]$, assuming $\epsilon \leq c^2$ (where $c = \Phi^{-1}(1-\delta)$) and $\sigma \geq 2$, the ND-DP-SGD algorithm satisfies (ϵ, δ)-DP for T iterations if:*

$$\sigma \geq \frac{q_s\sqrt{18(e-1)T\ln(1/\delta)}}{\epsilon}, \tag{12}$$

where $q_s = \frac{B_s}{n}$ is the Poisson subsampling rate.

Proof. See proof in Appendix G. □

Remark 5.2. The conditions $\epsilon \leq c^2$ and $\sigma \geq 2$ in Theorem 5.1 hold under standard privacy configurations. For instance, $\delta = 10^{-4}$ yields $c^2 \approx 13.83$, which is large enough to include practical privacy budget choices, such as $\epsilon = 1$.

Thanks to the automatic gradients reduction (Lemma 3.2) under interpolation, the variance of the stochastic noise automatically decreases along with the convergence of the loss function. Therefore, our analysis of the decaying injected noise naturally extends to the setting with additional stochastic noise, yielding the following theorem for empirical risk bound.

Theorem 5.3 (Empirical Risk of ND-DP-SGD on NNs)**.** *Suppose the optimization problem in Eq. (1) is in the interpolation regime, Assumption 3.1 holds, and μ-PL* and θ-Aiming conditions hold within a ball $\mathcal{B}_R(\mathbf{w}_0)$. For any $\delta_s, \delta_c \in (0,1)$ and given privacy parameters ϵ, δ, let $\alpha = \frac{\mu}{4}$, $\gamma = \eta(\theta - \beta\eta)\alpha$ and $K_{DP} = \sqrt{p} + \sqrt{2\ln\left(\frac{2T}{\delta_c}\right)}$. Set the hyperparameters of ND-DP-SGD as $\sigma = \frac{B_s\sqrt{18(e-1)T\ln(1/\delta)}}{n\epsilon}$, $B_s = \Omega(\ln T)$, $C_0 = \beta\sqrt{\frac{2\mathcal{L}(\mathbf{w}_0)}{\alpha}}$, $q = \sqrt{1-\frac{\gamma}{2}}$, and $\eta = \min\left\{\frac{\theta}{2\beta}, \frac{\theta\alpha B_s^2}{16\beta^2\sigma^2K_{DP}^2}\right\}$. If the ball radius $R \geq \frac{2\eta C_0}{1-q}\left(1 + K_{SGD} + \frac{\sigma K_{DP}}{B_s}\right)$ where K_{SGD} is the sampling error factor (see Lemma H.1), when the algorithm runs $T = \tilde{\Theta}(n^2\epsilon^2)$ steps, with probability at least $1 - \delta_c - \delta_s$, the empirical risk is bounded by:*

$$\mathcal{L}(\mathbf{w}_T) \leq \frac{\beta}{\alpha}\exp\left(-\tilde{\Theta}\left(\frac{n^2\epsilon^2}{p}\right)\right)\mathcal{L}(\mathbf{w}_0). \tag{13}$$

Proof. See proof in Appendix H. □

6 Conclusion

In this paper, we establish an exponential empirical risk bound for differentially private learning under interpolation. By leveraging the automatic gradients reduction phenomenon, we propose synchronously decaying the standard deviation of the injected privacy noise and the clipping threshold. This mechanism avoids noise accumulation and improves classical polynomial bounds to exponential $\mathcal{O}(\exp(-\tilde{\Theta}(n^2\epsilon^2/p)))$, for both strongly convex/PL problems and non-convex over-parameterized neural networks. It also extends to stochastic mini-batch training.

Limitations and future work. Our analysis is currently confined to the empirical risk. Extending these results to population risk remains outside the scope of this paper and is left for future work. For neural networks, our analysis builds on existing optimization theory for plain GD/SGD, which typically requires sufficiently large network width. Relaxing this width requirement is an interesting direction for future work, but would require corresponding advances in the existing optimization theory for plain GD/SGD.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing, grammar correction, and LaTeX formatting. We have not used generative AI tools for generating research ideas, formulating theoretical proofs, or designing algorithms. We have reviewed all AI-assisted work. We manually verified the polished text to ensure it accurately reflects our mathematical claims without altering the theoretical meaning. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This paper focuses on the theoretical analysis and algorithmic improvement of Differentially Private Gradient Descent (DP-GD). Our research aligns with the core principles of AI ethics by providing stronger mathematical guarantees for privacy-preserving machine learning. All empirical evaluations are conducted on standard public datasets (MNIST and CIFAR), and the study does not involve any sensitive human-subject data. We believe our findings will positively impact the secure deployment of machine learning models.

REPRODUCIBILITY STATEMENT

We have provided detailed proofs for all lemmas, propositions, and theorems in the appendices to ensure the reproducibility of our theoretical results. For the empirical results, all experimental details, including hyperparameter selections, dataset processing, and network architectures, are thoroughly documented in Appendix I.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International conference on machine learning*, pp. 242–252. PMLR, 2019.

Galen Andrew, Om Thakkar, Brendan McMahan, and Swaroop Ramaswamy. Differentially private learning with adaptive clipping. *Advances in neural information processing systems*, 34:17455–17466, 2021.

Raman Arora, Raef Bassily, Cristóbal Guzmán, Michael Menart, and Enayat Ullah. Differentially private generalized linear models revisited. *Advances in neural information processing systems*, 35:22505–22517, 2022.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science*, pp. 464–473. IEEE, 2014.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Simone Bombari and Marco Mondelli. Privacy for free in the overparameterized regime. *Proceedings of the National Academy of Sciences*, 122(15):e2423072122, 2025.

Zhiqi Bu, Jinshuo Dong, Qi Long, and Weijie J Su. Deep learning with gaussian differential privacy. *Harvard data science review*, 2020(23):10–1162, 2020.

Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. Automatic clipping: Differentially private deep learning made easier and stronger. *Advances in Neural Information Processing Systems*, 36:41727–41764, 2023.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and trends in Machine Learning*, 8(3-4):231–357, 2015.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

Xiangyi Chen, Steven Z Wu, and Mingyi Hong. Understanding gradient clipping in private sgd: A geometric perspective. *Advances in neural information processing systems*, 33:13773–13782, 2020.

Lynn Chua, Badih Ghazi, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Amer Sinha, and Chiyuan Zhang. Scalable dp-sgd: Shuffling vs. poisson subsampling. *Advances in Neural Information Processing Systems*, 37:70026–70047, 2024.

Yupeng Deng, Xiong Li, Jiabei He, Yuzhen Liu, and Wei Liang. A stochastic gradient descent algorithm based on adaptive differential privacy. In *International Conference on Collaborative Computing: Networking, Applications and Worksharing*, pp. 133–152. Springer, 2022.

Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):3–37, 2022.

Jian Du, Song Li, Xiangyi Chen, Siheng Chen, and Mingyi Hong. Dynamic differential-privacy preserving sgd. *arXiv preprint arXiv:2111.00173*, 2021.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pp. 1675–1685. PMLR, 2019a.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019b.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3-4):211–487, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Huang Fang, Xiaoyun Li, Chenglin Fan, and Ping Li. Improved convergence of differential private sgd with gradient clipping. In *The Eleventh International Conference on Learning Representations*, 2023.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 439–449, 2020.

Niv Haim, Gal Vardi, Gilad Yehudai, Ohad Shamir, and Michal Irani. Reconstructing training data from trained neural networks. *Advances in Neural Information Processing Systems*, 35: 22911–22924, 2022.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 795–811. Springer, 2016.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32, 2019.

Chaoyue Liu and Mikhail Belkin. Accelerating SGD with momentum for over-parameterized learning. In *8th International Conference on Learning Representations, ICLR 2020*, 2020.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59: 85–116, 2022.

Chaoyue Liu, Dmitriy Drusvyatskiy, Misha Belkin, Damek Davis, and Yian Ma. Aiming towards the minimizers: fast convergence of sgd for overparametrized problems. *Advances in neural information processing systems*, 36:60748–60767, 2023.

Yingchi Mao, Chenxin Li, Zibo Wang, Zijian Tu, and Ping Ping. Differential privacy in federated dynamic gradient clipping based on gradient norm. In *International Conference on Algorithms and Architectures for Parallel Processing*, pp. 24–41. Springer, 2023.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2013.

Samet Oymak and Mahdi Soltanolkotabi. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020.

Venkatadheeraj Pichapati, Ananda Theertha Suresh, Felix X Yu, Sashank J Reddi, and Sanjiv Kumar. Adaclip: Adaptive clipping for private sgd. *arXiv preprint arXiv:1908.07643*, 2019.

Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pp. 1679–1706, 1994.

Shuang Song, Kamalika Chaudhuri, Anand D Sarwate, et al. Stochastic gradient descent with differentially private updates. In *GlobalSIP*, pp. 245–248, 2013.

Shuang Song, Thomas Steinke, Om Thakkar, and Abhradeep Thakurta. Evading the curse of dimensionality in unconstrained private glms. In *International Conference on Artificial Intelligence and Statistics*, pp. 2638–2646. PMLR, 2021.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1195–1204. PMLR, 2019.

Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. *Advances in Neural Information Processing Systems*, 30, 2017.

Chengkun Wei, Weixian Li, Chen Gong, and Wenzhi Chen. Dc-sgd: Differentially private sgd with dynamic clipping through gradient norm distribution estimation. *IEEE Transactions on Information Forensics and Security*, 20:4498–4511, 2025.

Tianyu Xia, Shuheng Shen, Su Yao, Xinyi Fu, Ke Xu, Xiaolong Xu, and Xing Fu. Differentially private learning with per-sample adaptive clipping. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 10444–10452, 2023.

Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. Gradient perturbation is underrated for differentially private convex optimization. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pp. 3117–3123, 2020.

Lei Yu, Ling Liu, Calton Pu, Mehmet Emre Gursoy, and Stacey Truex. Differentially private model publishing for deep learning. In *2019 IEEE symposium on security and privacy (SP)*, pp. 332–349. IEEE, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Xinyue Zhang, Jiahao Ding, Maoqiang Wu, Stephen TC Wong, Hien Van Nguyen, and Miao Pan. Adaptive privacy preserving deep learning algorithms for medical data. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1169–1178, 2021.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109:467–492, 2020. doi: 10.1007/s10994-019-05839-6.

A Proof of Lemma 3.2

Proof. By the β-smoothness of the sample-wise loss function ℓ_i (Assumption 3.1), for any $\mathbf{w}, \mathbf{w}^+ \in \mathbb{R}^p$, we have the following inequality:

$$\ell_i(\mathbf{w}^+) \leq \ell_i(\mathbf{w}) + \langle \nabla \ell_i(\mathbf{w}), \mathbf{w}^+ - \mathbf{w} \rangle + \frac{\beta}{2} \|\mathbf{w}^+ - \mathbf{w}\|^2.$$

Set $\mathbf{w}^+ = \mathbf{w} - \eta \nabla \ell_i(\mathbf{w})$ for some step size $\eta > 0$. Substituting this $\mathbf{w}^+$ into the inequality yields:

$$\begin{aligned} \ell_i(\mathbf{w}^+) &\leq \ell_i(\mathbf{w}) - \eta \langle \nabla \ell_i(\mathbf{w}), \nabla \ell_i(\mathbf{w}) \rangle + \frac{\beta \eta^2}{2} \|\nabla \ell_i(\mathbf{w})\|^2 \\ &= \ell_i(\mathbf{w}) - \left(\eta - \frac{\beta \eta^2}{2} \right) \|\nabla \ell_i(\mathbf{w})\|^2. \end{aligned}$$

Maximizing the reduction with respect to η yields $\eta = \frac{1}{\beta}$. Substituting this back gives:

$$\ell_i(\mathbf{w}^+) \leq \ell_i(\mathbf{w}) - \frac{1}{2\beta} \|\nabla \ell_i(\mathbf{w})\|^2.$$

By the non-negativity of the loss, $\ell_i(\mathbf{w}^+) \geq 0$, which yields:

$$0 \leq \ell_i(\mathbf{w}) - \frac{1}{2\beta} \|\nabla \ell_i(\mathbf{w})\|^2.$$

Rearranging the terms, we obtain $\|\nabla \ell_i(\mathbf{w})\|^2 \leq 2\beta \ell_i(\mathbf{w})$ for each $i \in [n]$. Taking the expectation with respect to $i \in [n]$ on both sides, and noting that $\mathcal{L}(\mathbf{w}) = \mathbb{E}_{i \in [n]}[\ell_i(\mathbf{w})]$, we arrive at:

$$\mathbb{E}_{i \in [n]}[\|\nabla \ell_i(\mathbf{w})\|^2] \leq 2\beta \mathcal{L}(\mathbf{w}).$$

This completes the proof. □

B Noise Concentration Bound

Lemma B.1 (Noise Concentration). *Given any $\delta_c \in (0, 1)$, define the noise bound factor K as:*

$$K = \sigma \left(\sqrt{p} + \sqrt{2 \ln \left(\frac{2T}{\delta_c} \right)} \right). \tag{B.1}$$

For any $t \in [0, T-1]$, the ℓ_2-norm of the DP noise $\zeta^{(t)} \sim \mathcal{N}(0, C_t^2 \sigma^2 \boldsymbol{I}_p)$ in Algorithm 1 satisfies the following with probability at least $1 - \frac{\delta_c}{2T}$:

$$\|\zeta^{(t)}\| \leq K C_t.$$

Proof. According to Algorithm 1, the injected DP noise at step t follows a multivariate Gaussian distribution: $\zeta^{(t)} \sim \mathcal{N}(0, C_t^2 \sigma^2 \boldsymbol{I}_p)$. We can standardize this noise vector by defining $Z = \frac{\zeta^{(t)}}{C_t \sigma}$, which implies that $Z \sim \mathcal{N}(0, \boldsymbol{I}_p)$ is a standard p-dimensional Gaussian random vector.

Consider the L_2 norm function $f(Z) = \|Z\|$. By the triangle inequality, for any two vectors $x, y \in \mathbb{R}^p$, we have $|\|x\| - \|y\|| \leq \|x - y\|$. This demonstrates that $f(Z)$ is a 1-Lipschitz continuous function with respect to the Euclidean norm.

By the Gaussian Lipschitz concentration inequality (also known as the Borell-TIS inequality), the value of a 1-Lipschitz function of a standard Gaussian vector concentrates around its expectation. For any $u > 0$:

$$\mathbb{P}\Big(\|Z\| - \mathbb{E}[\|Z\|] \geq u \Big) \leq \exp\left(-\frac{u^2}{2} \right).$$

Apply Jensen's inequality to the concave square root function. Since the squared norm of a standard Gaussian vector follows a Chi-squared distribution with p degrees of freedom, $\mathbb{E}[\|Z\|^2] = p$, which yields:

$$\mathbb{E}[\|Z\|] = \mathbb{E}\left[\sqrt{\|Z\|^2} \right] \leq \sqrt{\mathbb{E}[\|Z\|^2]} = \sqrt{p}.$$

Substituting this upper bound into the concentration inequality:

$$\mathbb{P}\Big(\|Z\| \geq \sqrt{p} + u\Big) \leq \exp\left(-\frac{u^2}{2}\right).$$

Let $u = \sqrt{2\ln\left(\frac{2T}{\delta_c}\right)}$, then:

$$\mathbb{P}\Big(\|Z\| \geq \sqrt{p} + \sqrt{2\ln\left(\frac{2T}{\delta_c}\right)}\Big) \leq \exp\left(-\frac{2\ln\left(\frac{2T}{\delta_c}\right)}{2}\right) = \frac{\delta_c}{2T}.$$

Therefore, with probability at least $1 - \frac{\delta_c}{2T}$:

$$\|Z\| \leq \sqrt{p} + \sqrt{2\ln\left(\frac{2T}{\delta_c}\right)}.$$

Finally, by multiplying both sides of the inequality by $C_t\sigma$, we obtain the bound for $\zeta^{(t)}$:

$$\|\zeta^{(t)}\| \leq C_t\sigma\left(\sqrt{p} + \sqrt{2\ln\left(\frac{2T}{\delta_c}\right)}\right) = KC_t.$$

This completes the proof. □

C Proof of Theorem 3.6

C.1 Reduction to the PL Condition

By the μ-strong convexity of the empirical loss $\mathcal{L}$, for any $\mathbf{w}, \mathbf{w}'$, we have:

$$\mathcal{L}(\mathbf{w}') \geq \mathcal{L}(\mathbf{w}) + \langle \nabla\mathcal{L}(\mathbf{w}), \mathbf{w}' - \mathbf{w}\rangle + \frac{\mu}{2}\|\mathbf{w}' - \mathbf{w}\|^2.$$

We now take the minimum over all possible $\mathbf{w}'$ on both sides of this inequality. For the left-hand side, the minimum of $\mathcal{L}(\mathbf{w}')$ over the entire parameter space is achieved at the global optimum $\mathbf{w}^*$. Under the interpolation regime, this global minimum achieves zero loss, so $\min_{\mathbf{w}'} \mathcal{L}(\mathbf{w}') = \mathcal{L}(\mathbf{w}^*) = 0$. For the right-hand side, the expression is a quadratic function of the displacement, which attains its minimum when its derivative with respect to $\mathbf{w}'$ is zero, yielding $\mathbf{w}' - \mathbf{w} = -\frac{1}{\mu}\nabla\mathcal{L}(\mathbf{w})$. Evaluating the minimums of both sides gives:

$$\begin{aligned} 0 &\geq \min_{\mathbf{w}'}\left[\mathcal{L}(\mathbf{w}) + \langle \nabla\mathcal{L}(\mathbf{w}), \mathbf{w}' - \mathbf{w}\rangle + \frac{\mu}{2}\|\mathbf{w}' - \mathbf{w}\|^2\right] \\ &= \mathcal{L}(\mathbf{w}) - \frac{1}{2\mu}\|\nabla\mathcal{L}(\mathbf{w})\|^2. \end{aligned}$$

Rearranging this inequality establishes that standard strong convexity implies the global Polyak–Łojasiewicz (PL) condition:

$$\|\nabla\mathcal{L}(\mathbf{w})\|^2 \geq 2\mu\mathcal{L}(\mathbf{w}).$$

Under generalized strong convexity, the objective function is μ-strongly convex when restricted to the subspace orthogonal to the solution manifold $\mathcal{S}$. For any parameter $\mathbf{w}$, let $\mathbf{w}^* \in \mathcal{S}$ be its orthogonal projection onto the solution manifold. Both the displacement $\mathbf{w}^* - \mathbf{w}$ and the gradient $\nabla\mathcal{L}(\mathbf{w})$ lie within this orthogonal subspace. For any $\mathbf{w}'$ such that the displacement $\mathbf{w}' - \mathbf{w}$ is in the orthogonal subspace, the strong convexity lower bound holds:

$$\mathcal{L}(\mathbf{w}') \geq \mathcal{L}(\mathbf{w}) + \langle \nabla\mathcal{L}(\mathbf{w}), \mathbf{w}' - \mathbf{w}\rangle + \frac{\mu}{2}\|\mathbf{w}' - \mathbf{w}\|^2.$$

Since $\nabla\mathcal{L}(\mathbf{w})$ lies in the orthogonal subspace, we can take the minimum over all valid $\mathbf{w}'$ on both sides of this inequality. For the left-hand side, the minimum of $\mathcal{L}(\mathbf{w}')$ within this subspace is achieved at the projection point $\mathbf{w}^*$. Under the interpolation regime, this global minimum achieves zero loss, so $\min_{\mathbf{w}'} \mathcal{L}(\mathbf{w}') = \mathcal{L}(\mathbf{w}^*) = 0$. For the right-hand side, the quadratic expression attains its minimum

when its derivative with respect to $\mathbf{w}'$ is zero, which gives $\mathbf{w}' - \mathbf{w} = -\frac{1}{\mu}\nabla\mathcal{L}(\mathbf{w})$. Evaluating the minimums of both sides yields:

$$\begin{aligned} 0 &\geq \min_{\mathbf{w}'} \left[\mathcal{L}(\mathbf{w}) + \langle \nabla\mathcal{L}(\mathbf{w}), \mathbf{w}' - \mathbf{w}\rangle + \frac{\mu}{2}\|\mathbf{w}' - \mathbf{w}\|^2\right] \\ &= \mathcal{L}(\mathbf{w}) - \frac{1}{2\mu}\|\nabla\mathcal{L}(\mathbf{w})\|^2. \end{aligned}$$

Rearranging this inequality yields the global PL condition:

$$\|\nabla\mathcal{L}(\mathbf{w})\|^2 \geq 2\mu\mathcal{L}(\mathbf{w}).$$

Thus, standard strong convexity and generalized strong convexity both imply the global PL condition.

Furthermore, as established by Karimi et al. (2016), the PL condition implies the Quadratic Growth (QG) property with respect to the optimal set $\mathcal{S}$:

$$\operatorname{dist}(\mathbf{w}, \mathcal{S}) \leq \sqrt{\frac{2}{\mu}\mathcal{L}(\mathbf{w})}.$$

The subsequent mathematical induction relies solely on the PL and QG properties. This proof framework holds for any objective function satisfying standard strong convexity, generalized strong convexity, or the global PL condition.

C.2 Inductive Proof of Exponential Convergence

Proof. With K defined in equation B.1, let $\mathcal{E}_1$ be the event that the DP noise concentration bound holds for all $t \in [0, T-1]$: $\|\zeta^{(t)}\| \leq KC_t$. By Lemma B.1, $\mathbb{P}(\mathcal{E}_1) \geq 1 - \frac{\delta_c}{2}$. Additionally, let $\mathcal{E}_2$ be the event that the Gaussian cross-term projection bound holds for all $t \in [0, T-1]$. As we will show, the projection of the isotropic noise $\zeta^{(t)}$ onto the fixed gradient direction at step t admits a failure probability of $\frac{\delta_c}{2T}$ per step. By the union bound, $\mathbb{P}(\mathcal{E}_2) \geq 1 - \frac{\delta_c}{2}$. We define the joint global success event as $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2$. By the union bound, $\mathbb{P}(\mathcal{E}) \geq 1 - \delta_c$. Conditioned on $\mathcal{E}$, all subsequent bounds hold, and we prove by mathematical induction that for all $t \in [0, T]$:

$$\mathcal{L}(\mathbf{w}_t) \leq \frac{\mu}{2\beta^2}C_t^2. \tag{C.1}$$

Base Case ($t = 0$): Substituting C_0 yields $\mathcal{L}(\mathbf{w}_0) \leq \frac{\mu}{2\beta^2}\left(\beta\sqrt{\frac{2}{\mu}\mathcal{L}(\mathbf{w}_0)}\right)^2 = \mathcal{L}(\mathbf{w}_0)$, which holds trivially.

Inductive Step ($t \to t+1$): Assume the hypothesis holds at step t. We evaluate the loss at step $t+1$. Under the condition $q < 1$, the sequence C_t is decreasing. From the QG property and the inductive hypothesis, the distance to the optimal set is bounded by:

$$\operatorname{dist}(\mathbf{w}_t, \mathcal{S}) \leq \sqrt{\frac{2}{\mu}\mathcal{L}(\mathbf{w}_t)} \leq \frac{C_t}{\beta}.$$

Applying the β-smoothness of each ℓ_i and the interpolation condition $\nabla\ell_i(\mathbf{w}^*) = \mathbf{0}$, we have $\max_{i\in[n]}\|\nabla\ell_i(\mathbf{w}_t)\| = \max_{i\in[n]}\|\nabla\ell_i(\mathbf{w}_t) - \nabla\ell_i(\mathbf{w}^*)\| \leq \beta\|\mathbf{w}_t - \mathbf{w}^*\| = \beta\operatorname{dist}(\mathbf{w}_t, \mathcal{S}) \leq C_t$. Consequently, the clipping step is bypassed, and the parameter update reduces to $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\nabla\mathcal{L}(\mathbf{w}_t) - \frac{\eta}{n}\zeta^{(t)}$. By the β-smoothness of $\mathcal{L}$:

$$\begin{aligned} \mathcal{L}(\mathbf{w}_{t+1}) &\leq \mathcal{L}(\mathbf{w}_t) + \langle \nabla\mathcal{L}(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle + \frac{\beta}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 \\ &= \mathcal{L}(\mathbf{w}_t) + \left\langle \nabla\mathcal{L}(\mathbf{w}_t), -\eta\nabla\mathcal{L}(\mathbf{w}_t) - \frac{\eta}{n}\zeta^{(t)}\right\rangle + \frac{\beta}{2}\left\|\eta\nabla\mathcal{L}(\mathbf{w}_t) + \frac{\eta}{n}\zeta^{(t)}\right\|^2 \\ &= \mathcal{L}(\mathbf{w}_t) - \eta\left(1 - \frac{\beta\eta}{2}\right)\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2 - \frac{\eta}{n}(1 - \beta\eta)\langle\nabla\mathcal{L}(\mathbf{w}_t), \zeta^{(t)}\rangle + \frac{\beta\eta^2}{2n^2}\|\zeta^{(t)}\|^2. \end{aligned}$$

Since the learning rate $\eta \leq \frac{1}{\beta}$, we have $1 - \frac{\beta\eta}{2} \geq \frac{1}{2}$ and $0 \leq 1 - \beta\eta < 1$. Applying the PL condition $\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2 \geq 2\mu\mathcal{L}(\mathbf{w}_t)$ to the negative term, and bounding the squared noise term as $\|\zeta^{(t)}\|^2 \leq K^2C_t^2$:

$$\mathcal{L}(\mathbf{w}_{t+1}) \leq (1 - \mu\eta)\mathcal{L}(\mathbf{w}_t) + \frac{\beta\eta^2}{2n^2}K^2C_t^2 - \frac{\eta}{n}(1 - \beta\eta)\langle\nabla\mathcal{L}(\mathbf{w}_t), \zeta^{(t)}\rangle.$$

The Gaussian Cross Term: Conditioned on the state at step t, the gradient $\nabla\mathcal{L}(\mathbf{w}_t)$ is fixed. Since the injected noise $\zeta^{(t)}$ is drawn from an isotropic Gaussian $\mathcal{N}(0, C_t^2\sigma^2\boldsymbol{I}_p)$, its projection forms a one-dimensional zero-mean Gaussian random variable with variance $\nu^2 = \|\nabla\mathcal{L}(\mathbf{w}_t)\|^2 C_t^2\sigma^2$. Setting the two-sided Gaussian tail bound $\mathbb{P}(|X| > \tau) \le 2\exp(-\tau^2/(2\nu^2))$ to a failure probability of $\delta_c/(2T)$ yields the threshold $\tau = \nu\sqrt{2\ln(4T/\delta_c)}$. Conditioned on the global success event $\mathcal{E}$, this cross-term bound $\mathcal{E}_2$ holds for step t, giving:

$$-\frac{\eta(1-\beta\eta)}{n}\langle\nabla\mathcal{L}(\mathbf{w}_t), \zeta^{(t)}\rangle \le \frac{\eta(1-\beta\eta)}{n}\|\nabla\mathcal{L}(\mathbf{w}_t)\| C_t\sigma\sqrt{2\ln(4T/\delta_c)}.$$

Substituting the bound $\|\nabla\mathcal{L}(\mathbf{w}_t)\| \le \sqrt{2\beta\mathcal{L}(\mathbf{w}_t)}$ via Lemma 3.2, and applying the inductive hypothesis $\sqrt{\mathcal{L}(\mathbf{w}_t)} \le \frac{\sqrt{\mu}}{\sqrt{2\beta}}C_t$, we expand the bound:

$$\begin{aligned}
\mathcal{L}(\mathbf{w}_{t+1}) &\le (1-\mu\eta)\mathcal{L}(\mathbf{w}_t) + \frac{\beta\eta^2K^2}{2n^2}C_t^2 + \frac{\eta(1-\beta\eta)\sqrt{2\beta}}{n}\sqrt{\mathcal{L}(\mathbf{w}_t)}C_t\sigma\sqrt{2\ln(4T/\delta_c)} \\
&\le (1-\mu\eta)\frac{\mu}{2\beta^2}C_t^2 + \frac{\beta\eta^2K^2}{2n^2}C_t^2 + \frac{\eta(1-\beta\eta)\sqrt{\mu}}{n\sqrt{\beta}}C_t^2\sigma\sqrt{2\ln(4T/\delta_c)} \\
&= \left((1-\mu\eta) + \frac{\beta^3\eta^2K^2}{\mu n^2} + \frac{2\eta(1-\beta\eta)\beta^{3/2}}{\sqrt{\mu}n}\sigma\sqrt{2\ln(4T/\delta_c)}\right)\frac{\mu}{2\beta^2}C_t^2 \\
&\le q^2\frac{\mu}{2\beta^2}C_t^2 = \frac{\mu}{2\beta^2}C_{t+1}^2.
\end{aligned}$$

This proves the inductive bound (C.1) for step $t+1$, provided we set the squared decay rate q^2 to upper bound the bracketed expression.

Final Convergence Evaluation:

Convergence with $q^2 = 1 - \frac{\mu\eta}{2}$ requires the combined noise components to satisfy:

$$\frac{\beta^3\eta^2K^2}{\mu n^2} + \frac{2\eta(1-\beta\eta)\beta^{3/2}}{\sqrt{\mu}n}\sigma\sqrt{2\ln(4T/\delta_c)} \le \frac{\mu\eta}{2}.$$

Allocating the $\frac{\mu\eta}{2}$ budget equally to both terms decouples the constraints. For the squared noise term, substituting $K = \tilde{\mathcal{O}}\left(\frac{\sqrt{pT}}{\epsilon}\right)$ yields:

$$\frac{\beta^3\eta^2K^2}{\mu n^2} \le \frac{\mu\eta}{4} \implies \eta \le \tilde{\mathcal{O}}\left(\frac{\mu^2n^2\epsilon^2}{\beta^3pT}\right).$$

For the cross term, applying $\sigma\sqrt{2\ln(4T/\delta_c)} = \tilde{\mathcal{O}}\left(\frac{\sqrt{T}}{\epsilon}\right)$ yields:

$$\frac{2\eta(1-\beta\eta)\beta^{3/2}}{\sqrt{\mu}n}\sigma\sqrt{2\ln(4T/\delta_c)} \le \frac{\mu\eta}{4} \implies T \le \tilde{\mathcal{O}}(n^2\epsilon^2).$$

To satisfy all of the above constraints simultaneously, we set the learning rate as $\eta = \min\left\{\frac{1}{\beta}, \frac{\mu^2n^2}{4\beta^3K^2}\right\}$ and the total iterations as $T = \tilde{\Theta}(n^2\epsilon^2)$, which matches the hyperparameter configuration defined in Theorem 3.6.

Using the inequality $(1-x)^k \le e^{-xk}$ for $x \in (0,1)$, we bound the geometric decay term by $q^{2T} \le \exp\left(-\frac{\mu\eta}{2}T\right)$. Substituting the hyperparameter values specified in Theorem 3.6 (yielding $\eta = \tilde{\Theta}\left(\frac{n^2\epsilon^2}{pT}\right)$ and $T = \tilde{\Theta}(n^2\epsilon^2)$) into the decay term, and applying the inductive bound (C.1) at step T, the final empirical risk is bounded by:

$$\mathcal{L}(\mathbf{w}_T) \le \frac{\mu}{2\beta^2}C_T^2 = \mathcal{L}(\mathbf{w}_0)q^{2T} \le \exp\left(-\tilde{\Theta}\left(\frac{n^2\epsilon^2}{p}\right)\right)\mathcal{L}(\mathbf{w}_0).$$

□

D PROOF OF THEOREM 3.7

In this section, we provide the detailed proof for Theorem 3.7. The privacy guarantee is established using the Moments Accountant framework (Abadi et al., 2016).

D.1 MOMENTS ACCOUNTANT

The Moments Accountant is a refined privacy accounting framework that provides a tighter privacy guarantee than standard composition theorems by tracking the higher-order moments of the privacy loss distribution.

Let $\mathcal{M}$ be a randomized mechanism. For adjacent datasets $\mathcal{D}, \mathcal{D}'$ and auxiliary input aux, the privacy loss at outcome $\mathbf{w}_{t+1}$ is defined as:

$$c(\mathbf{w}_{t+1}; \mathcal{M}, aux, \mathcal{D}, \mathcal{D}') = \ln \frac{\mathbb{P}(\mathcal{M}(aux, \mathcal{D}) = \mathbf{w}_{t+1})}{\mathbb{P}(\mathcal{M}(aux, \mathcal{D}') = \mathbf{w}_{t+1})}.$$

We define the λ-th moment as the log moment generating function of the privacy loss:

$$\alpha_{\mathcal{M}}(\lambda; aux, \mathcal{D}, \mathcal{D}') = \ln \mathbb{E}_{\mathbf{w}_{t+1} \sim \mathcal{M}(aux, \mathcal{D})}[\exp(\lambda c(\mathbf{w}_{t+1}; \mathcal{M}, aux, \mathcal{D}, \mathcal{D}'))].$$

Then we define:

$$\alpha_{\mathcal{M}}(\lambda) = \sup_{aux, \mathcal{D}, \mathcal{D}'} \alpha_{\mathcal{M}}(\lambda; aux, \mathcal{D}, \mathcal{D}').$$

According to Theorem 2 in Abadi et al. (2016), the moments satisfy two key properties:

1. **Composability**: For a sequence of adaptive mechanisms $\mathcal{M} = (\mathcal{M}_0, \ldots, \mathcal{M}_{T-1})$, the total moment is bounded by $\alpha_{\mathcal{M}}(\lambda) \leq \sum_{t=0}^{T-1} \alpha_{\mathcal{M}_t}(\lambda)$.
2. **Tail Bound**: For any $\epsilon > 0$, the mechanism $\mathcal{M}$ is (ϵ, δ)-DP if $\delta \geq \min_{\lambda > 0} \exp(\alpha_{\mathcal{M}}(\lambda) - \lambda\epsilon)$.

D.2 SENSITIVITY ANALYSIS OF THE UPDATE

Let $\mathcal{M}_t$ denote the update mechanism at iteration t. The update rule in Algorithm 1 is:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{\eta}{n} \sum_{i=1}^{n} \tilde{g}_i^{(t)} - \frac{\eta}{n} \zeta^{(t)},$$

where $\zeta^{(t)} \sim \mathcal{N}(0, C_t^2 \sigma^2 \boldsymbol{I}_p)$. Conditioned on the previous iterate $\mathbf{w}_t$ and the clipping threshold C_t, $\mathbf{w}_{t+1}$ follows a multivariate Gaussian distribution:

$$\mathbf{w}_{t+1} \sim \mathcal{N}\left(\mu_{t,\mathcal{D}}, \xi_t^2 \boldsymbol{I}_p\right),$$

where the mean $\mu_{t,\mathcal{D}} = \mathbf{w}_t - \frac{\eta}{n} \sum_{i \in \mathcal{D}} \tilde{g}_i^{(t)}$ and the standard deviation $\xi_t = \frac{\eta C_t \sigma}{n}$.

Lemma D.1 (Sensitivity of the Update Mean)**.** *Consider two adjacent datasets $\mathcal{D}$ and $\mathcal{D}'$ differing by one sample. The L_2 sensitivity of the mean $\mu_{t,\mathcal{D}}$ is bounded by:*

$$\sup_{\mathcal{D}, \mathcal{D}'} \|\mu_{t,\mathcal{D}} - \mu_{t,\mathcal{D}'}\| \leq \frac{2\eta C_t}{n}.$$

Proof. Let the datasets $\mathcal{D}$ and $\mathcal{D}'$ differ by replacing the j-th sample with sample j'. The means of the update distributions differ only in the gradient sum term:

$$\begin{aligned}
\|\mu_{t,\mathcal{D}} - \mu_{t,\mathcal{D}'}\| &= \left\| \left(\mathbf{w}_t - \frac{\eta}{n} \sum_{i \in \mathcal{D}} \tilde{g}_i^{(t)} \right) - \left(\mathbf{w}_t - \frac{\eta}{n} \sum_{i \in \mathcal{D}'} \tilde{g}_i^{(t)} \right) \right\| \\
&= \left\| -\frac{\eta}{n} (\tilde{g}_j^{(t)} - \tilde{g}_{j'}^{(t)}) \right\| \\
&= \frac{\eta}{n} \|\tilde{g}_j^{(t)} - \tilde{g}_{j'}^{(t)}\| \\
&\leq \frac{\eta}{n} (\|\tilde{g}_j^{(t)}\| + \|\tilde{g}_{j'}^{(t)}\|).
\end{aligned}$$

By the definition of the clipping operation in Algorithm 1, any clipped gradient satisfies $\|\tilde{g}_i^{(t)}\| \leq C_t$. Therefore:

$$\|\mu_{t,\mathcal{D}} - \mu_{t,\mathcal{D}'}\| \leq \frac{\eta}{n}(C_t + C_t) = \frac{2\eta C_t}{n}.$$

Taking the supremum yields the proof. □

D.3 Single-Step Moment Bound

Next, we bound the log-moment for a single Gaussian mechanism step.

Lemma D.2 (log-moment of $\mathcal{M}_t$)**.** *The log-moment of the privacy loss for the mechanism $\mathcal{M}_t$ at iteration t satisfies:*

$$\alpha_{\mathcal{M}_t}(\lambda) \leq \frac{2}{\sigma^2}(\lambda^2 + \lambda).$$

Proof. Let $P \sim \mathcal{N}(\mu_{t,\mathcal{D}}, \xi_t^2 \boldsymbol{I}_p)$ and $Q \sim \mathcal{N}(\mu_{t,\mathcal{D}'}, \xi_t^2 \boldsymbol{I}_p)$. Let $\Delta_t = \mu_{t,\mathcal{D}} - \mu_{t,\mathcal{D}'}$. The privacy loss variable $c(\mathbf{w}_{t+1}) = \ln(P(\mathbf{w}_{t+1})/Q(\mathbf{w}_{t+1}))$ evaluated at output $\mathbf{w}_{t+1}$ is given by:

$$\begin{aligned} c(\mathbf{w}_{t+1}) &= \ln\left(\frac{\exp(-\|\mathbf{w}_{t+1} - \mu_{t,\mathcal{D}}\|^2/2\xi_t^2)}{\exp(-\|\mathbf{w}_{t+1} - \mu_{t,\mathcal{D}'}\|^2/2\xi_t^2)}\right) \\ &= \frac{1}{2\xi_t^2}\left(\|\mathbf{w}_{t+1} - \mu_{t,\mathcal{D}'}\|^2 - \|\mathbf{w}_{t+1} - \mu_{t,\mathcal{D}}\|^2\right) \\ &= \frac{1}{2\xi_t^2}\left(2(\mathbf{w}_{t+1})^T(\mu_{t,\mathcal{D}} - \mu_{t,\mathcal{D}'}) + \|\mu_{t,\mathcal{D}'}\|^2 - \|\mu_{t,\mathcal{D}}\|^2\right). \end{aligned}$$

Let $\Delta_t = \mu_{t,\mathcal{D}} - \mu_{t,\mathcal{D}'}$. We can rewrite the output $\mathbf{w}_{t+1}$ as $\mathbf{w}_{t+1} = \mu_{t,\mathcal{D}} + z$, where $z \sim \mathcal{N}(0, \xi_t^2 \boldsymbol{I}_p)$ is the noise vector. Substituting this back:

$$\begin{aligned} c(\mathbf{w}_{t+1}) &= \frac{1}{2\xi_t^2}\left(2(\mu_{t,\mathcal{D}} + z)^T\Delta_t + \|\mu_{t,\mathcal{D}} - \Delta_t\|^2 - \|\mu_{t,\mathcal{D}}\|^2\right) \\ &= \frac{1}{2\xi_t^2}\left(2\mu_{t,\mathcal{D}}^T\Delta_t + 2z^T\Delta_t + \|\mu_{t,\mathcal{D}}\|^2 - 2\mu_{t,\mathcal{D}}^T\Delta_t + \|\Delta_t\|^2 - \|\mu_{t,\mathcal{D}}\|^2\right) \\ &= \frac{1}{2\xi_t^2}\left(2z^T\Delta_t + \|\Delta_t\|^2\right) \\ &= \frac{1}{\xi_t^2}z^T\Delta_t + \frac{\|\Delta_t\|^2}{2\xi_t^2}. \end{aligned}$$

The λ-th moment is the ln-MGF of $c(\mathbf{w}_{t+1})$:

$$\alpha_{\mathcal{M}_t}(\lambda) = \ln \mathbb{E}_{z\sim\mathcal{N}(0,\xi_t^2\boldsymbol{I}_p)}\left[\exp(\lambda c(\mathbf{w}_{t+1}))\right] = \ln \mathbb{E}\left[\exp\left(\frac{\lambda}{\xi_t^2}z^T\Delta_t + \frac{\lambda\|\Delta_t\|^2}{2\xi_t^2}\right)\right].$$

The term $z^T\Delta_t$ is a linear combination of Gaussian variables, so $z^T\Delta_t \sim \mathcal{N}(0, \xi_t^2\|\Delta_t\|^2)$. Using the property that for a variable $X \sim \mathcal{N}(0, \tau^2)$, $\mathbb{E}[e^{kX}] = e^{k^2\tau^2/2}$, we have:

$$\begin{aligned} \alpha_{\mathcal{M}_t}(\lambda) &= \ln\exp\left(\frac{1}{2}\left(\frac{\lambda}{\xi_t^2}\right)^2\xi_t^2\|\Delta_t\|^2\right) + \frac{\lambda\|\Delta_t\|^2}{2\xi_t^2} \\ &= \frac{\lambda^2\|\Delta_t\|^2}{2\xi_t^2} + \frac{\lambda\|\Delta_t\|^2}{2\xi_t^2} \\ &= \frac{\|\Delta_t\|^2}{2\xi_t^2}(\lambda^2 + \lambda). \end{aligned}$$

From Lemma D.1, we know $\|\Delta_t\| \leq \frac{2\eta C_t}{n}$. The variance is $\xi_t^2 = \left(\frac{\eta C_t \sigma}{n}\right)^2$. Substituting these values:

$$\frac{\|\Delta_t\|^2}{2\xi_t^2} \leq \frac{(2\eta C_t/n)^2}{2(\eta C_t\sigma/n)^2} = \frac{4\eta^2 C_t^2/n^2}{2\eta^2 C_t^2\sigma^2/n^2} = \frac{2}{\sigma^2}.$$

Thus, the log-moment is bounded by:

$$\alpha_{\mathcal{M}_t}(\lambda) \leq \frac{2}{\sigma^2}(\lambda^2 + \lambda).$$

□

D.4 Proof of Theorem 3.7

Proof. Recall from the single-step analysis that the log-moment for the t-th iteration is bounded by:

$$\alpha_{\mathcal{M}_t}(\lambda) \leq \frac{2}{\sigma^2}(\lambda^2 + \lambda).$$

Using the composability property of the Moments Accountant method, the total privacy loss $\alpha_{\mathcal{M}}(\lambda)$ is bounded by the sum of the moments:

$$\alpha_{\mathcal{M}}(\lambda) \leq \sum_{t=0}^{T-1} \alpha_{\mathcal{M}_t}(\lambda) \leq \frac{2T}{\sigma^2}(\lambda^2 + \lambda).$$

Set $\sigma \geq \frac{\sqrt{32T\ln(1/\delta)}}{\epsilon}$. Rearranging the inequality yields:

$$\frac{2T}{\sigma^2} \leq \frac{\epsilon^2}{16\ln(1/\delta)}.$$

Then, considering $\lambda > 0$, $\delta \in (0,1)$, and $\epsilon \in (0, 8\ln(1/\delta))$, we have that

$$\begin{aligned}
\exp\left(\alpha_{\mathcal{M}}(\lambda) - \lambda\epsilon\right) &\leq \exp\left(\frac{\epsilon^2}{16\ln(1/\delta)}(\lambda + \lambda^2) - \lambda\epsilon\right) \\
&= \exp\left(\frac{\epsilon^2}{16\ln(1/\delta)}\lambda^2 - \left(1 - \frac{\epsilon}{16\ln(1/\delta)}\right)\lambda\epsilon\right) \\
&\leq \exp\left(\frac{\epsilon^2}{16\ln(1/\delta)}\lambda^2 - \frac{\lambda\epsilon}{2}\right).
\end{aligned}$$

Minimizing the upper bound on the right-hand side by setting $\lambda^* = 4\ln(1/\delta)/\epsilon$ yields:

$$\begin{aligned}
\exp\left(\alpha_{\mathcal{M}}(\lambda^*) - \lambda^*\epsilon\right) &\leq \exp\left(\frac{\epsilon^2}{16\ln(1/\delta)}\lambda^{*2} - \frac{\lambda^*\epsilon}{2}\right) \\
&= \exp\left(\ln(1/\delta) - 2\ln(1/\delta)\right) \\
&= \exp\left(-\ln(1/\delta)\right) \\
&= \delta.
\end{aligned}$$

Then, by the tail bound from Theorem 2 in Abadi et al. (2016), we have that $\mathcal{M}$ is (ϵ, δ)-differentially private. □

E Analysis with Gaussian Differential Privacy

In this section, we analyze the full-batch ND-DP-GD algorithm using the f-Differential Privacy (f-DP) framework (Dong et al., 2022). Specifically, we use Gaussian Differential Privacy (GDP), a family within f-DP, to analyze the composition of Gaussian mechanisms. This analysis allows for a smaller noise multiplier than the standard Moments Accountant (Theorem 3.7), and establishes the base mechanism for the subsampled ND-DP-SGD privacy proof in Appendix G.

In GDP, the privacy leakage of a Gaussian mechanism is characterized by the ratio of the sensitivity and the noise standard deviation, denoted as μ (Dong et al., 2022, Theorem 1). For our ND-DP-GD algorithm, at iteration t, the L_2 sensitivity of the mean update is $\Delta_t = \frac{2\eta C_t}{n}$, and the injected Gaussian noise standard deviation is $\xi_t = \frac{\eta C_t \sigma}{n}$.

The GDP parameter μ_t for a single step is:

$$\mu_t = \frac{\Delta_t}{\xi_t} = \frac{2}{\sigma}$$

According to the composition theorem of GDP (Dong et al., 2022, Corollary 2), the total privacy cost after T iterations is:

$$\mu_{total} = \sqrt{\sum_{t=1}^{T} \mu_t^2} = \frac{2\sqrt{T}}{\sigma}$$

GDP provides the formula to translate μ_{total} back to the standard (ϵ, δ)-DP definition (Dong et al., 2022, Corollary 1):

$$\delta(\epsilon, \mu) = \Phi\left(-\frac{\epsilon}{\mu} + \frac{\mu}{2}\right) - e^{\epsilon}\Phi\left(-\frac{\epsilon}{\mu} - \frac{\mu}{2}\right)$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution. Since $e^{\epsilon}\Phi(-\frac{\epsilon}{\mu} - \frac{\mu}{2}) > 0$, we can establish an upper bound by:

$$\delta(\epsilon, \mu_{total}) < \Phi\left(-\frac{\epsilon}{\mu_{total}} + \frac{\mu_{total}}{2}\right)$$

We restrict this upper bound to not exceed our target δ. Let $c = \Phi^{-1}(1-\delta)$. Due to symmetry, $\Phi^{-1}(\delta) = -c$. This gives:

$$-\frac{\epsilon}{\mu_{total}} + \frac{\mu_{total}}{2} \leq -c$$

Multiplying by $2\mu_{total}$ and solving for the positive root, we obtain the closed-form upper bound for μ_{total}:

$$\begin{aligned} \mu_{total}^2 + 2c\mu_{total} - 2\epsilon &\leq 0 \\ \mu_{total} &\leq \sqrt{c^2 + 2\epsilon} - c \end{aligned} \tag{E.1}$$

Substitute $\mu_{total} = \frac{2\sqrt{T}}{\sigma}$ and solve for σ:

$$\sigma \geq \frac{\sqrt{T}(\sqrt{c^2 + 2\epsilon} + c)}{\epsilon}$$

For any $c > 0$ and $\epsilon > 0$, $(c + \frac{\epsilon}{c})^2 = c^2 + 2\epsilon + \frac{\epsilon^2}{c^2} > c^2 + 2\epsilon$. Therefore, we have $\sqrt{c^2 + 2\epsilon} < c + \frac{\epsilon}{c}$. Under the condition $\epsilon \leq c^2$ (e.g., $c^2 \approx 13.83$ for $\delta = 10^{-4}$), we have $\frac{\epsilon}{c} \leq c$. Substituting these into the noise lower bound yields:

$$\frac{\sqrt{T}(\sqrt{c^2 + 2\epsilon} + c)}{\epsilon} < \frac{\sqrt{T}\left(2c + \frac{\epsilon}{c}\right)}{\epsilon} \leq \frac{3c\sqrt{T}}{\epsilon}$$

For $\delta \leq 0.5$, we have $c = \Phi^{-1}(1-\delta) \geq 0$. By the Chernoff bound for the standard normal distribution, the tail probability is bounded by $\mathbb{P}(Z > c) \leq e^{-c^2/2}$. By the definition of the CDF, $\mathbb{P}(Z > c) = 1 - \Phi(c) = \delta$. Thus, we have $\delta \leq e^{-c^2/2}$, which yields the upper bound $c \leq \sqrt{2\ln(1/\delta)}$. Then we have:

$$\frac{\sqrt{T}(\sqrt{c^2 + 2\epsilon} + c)}{\epsilon} < \frac{\sqrt{18T\ln(1/\delta)}}{\epsilon}$$

Thus, the requirement for σ is:

$$\sigma \geq \frac{\sqrt{18T\ln(1/\delta)}}{\epsilon}$$

Compared to the noise requirement $\sigma \geq \frac{\sqrt{32T\ln(1/\delta)}}{\epsilon}$ derived via the Moments Accountant, the GDP framework can achieve a constant-level improvement, while still preserving the $\mathcal{O}(\exp(-\tilde{\Theta}(n^2\epsilon^2/p)))$ empirical risk bound.

F Proof of Theorem 4.4

We first restate a useful property that follows from the PL* condition. This quadratic growth property bounds the distance to the solution manifold $\mathcal{S}$.

Proposition F.1 (Quadratic Growth, Liu et al. (2023)). *Suppose the μ-PL* condition holds in $\mathcal{B}_R(\mathbf{w}_0)$ and $\mathcal{L}(\mathbf{w}_0) \leq \frac{1}{2}\mu R^2$. Then $\mathcal{S} \cap \mathcal{B}_{R/2}(\mathbf{w}_0) \neq \varnothing$, and for all $\mathbf{w} \in \mathcal{B}_{R/2}(\mathbf{w}_0)$, we have $\mathcal{L}(\mathbf{w}) \geq \frac{\mu}{8}\operatorname{dist}^2(\mathbf{w}, \mathcal{S}) = \frac{\alpha}{2}\operatorname{dist}^2(\mathbf{w}, \mathcal{S})$.*

F.1 Hyperparameter Calibration

Lemma F.2. *Provided the hyperparameters are calibrated as in Theorem 4.4, and with K defined in equation B.1, the decay rate $q = \sqrt{1 - \frac{\gamma}{2}}$ satisfies:*

$$(1-\gamma) + \left(\frac{\eta K \beta}{n}\right)^2 + \frac{2\eta\sigma\beta}{n}\sqrt{1-\gamma}\sqrt{2\ln(4T/\delta_c)} \le q^2 = 1 - \frac{\gamma}{2}.$$

Proof. Subtracting $(1-\gamma)$ and allocating the capacity $\frac{\gamma}{2}$ equally, we require each term to be bounded by $\frac{\gamma}{4}$.

For the first term, substituting $\gamma = \eta(\theta - \beta\eta)\alpha$ and $K = \tilde{\mathcal{O}}\left(\frac{\sqrt{Tp}}{\epsilon}\right)$:

$$\left(\frac{\eta K \beta}{n}\right)^2 \le \frac{\gamma}{4} \implies \eta \le \tilde{\mathcal{O}}\left(\frac{n^2\epsilon^2}{pT}\right).$$

For the second term, substituting $\gamma = \eta(\theta - \beta\eta)\alpha$ and $\sigma\sqrt{2\ln(4T/\delta_c)} = \tilde{\mathcal{O}}\left(\frac{\sqrt{T}}{\epsilon}\right)$:

$$\frac{2\eta\sigma\beta}{n}\sqrt{1-\gamma}\sqrt{2\ln(4T/\delta_c)} \le \frac{\gamma}{4} \implies T \le \tilde{\mathcal{O}}\left(n^2\epsilon^2\right).$$

To satisfy these bounds simultaneously, we set the learning rate as $\eta = \min\left\{\frac{\theta}{2\beta}, \frac{\theta\alpha n^2}{8\beta^2 K^2}\right\}$ and the total iterations as $T = \tilde{\Theta}(n^2\epsilon^2)$, which matches the hyperparameter configuration defined in Theorem 4.4. □

F.2 No-Clipping and Bounded Trajectory

Now we construct a joint induction to establish the "No-Clipping" and "Non-Escape" properties simultaneously.

Proposition F.3. *Under the same conditions as Theorem 4.4, with probability at least $1 - \delta_c$, the following statements hold simultaneously for all $t \in [0, T]$:*

1. $\max_{i\in[n]} \|\nabla\ell_i(\mathbf{w}_t)\| \le C_t$.
2. $\mathbf{w}_t \in \mathcal{B}_{R/2}(\mathbf{w}_0)$.

Proof. Let $\mathcal{E}_1$ be the event where the DP noise bound (Lemma B.1) holds for all $t \in [0, T-1]$ (with probability $\ge 1 - \frac{\delta_c}{2}$). Additionally, let $\mathcal{E}_2$ be the event that the Gaussian cross-term projection bound holds for all $t \in [0, T-1]$. As we will show, the projection of the isotropic noise $\zeta^{(t)}$ onto the fixed direction at step t admits a failure probability of $\frac{\delta_c}{2T}$ per step. By the union bound, $\mathbb{P}(\mathcal{E}_2) \ge 1 - \frac{\delta_c}{2}$. We define the joint global success event as $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2$, which guarantees $\mathbb{P}(\mathcal{E}) \ge 1 - \delta_c$.

We define $U_t \triangleq \|\mathbf{w}_t - \bar{\mathbf{w}}_t\|^2$, and let the local region radius be bounded by $R = 2\eta\left(1 + \frac{K}{n}\right)\frac{C_0}{1-q}$. Conditioned on $\mathcal{E}$, we prove by mathematical induction that for all $t \in [0, T]$, the following two statements hold simultaneously: (1) $\sqrt{U_t} \le \frac{C_t}{\beta}$. (2) $\mathbf{w}_t \in \mathcal{B}_{R/2}(\mathbf{w}_0)$.

Base Case ($t = 0$): Proposition F.1 guarantees that the Quadratic Growth condition $\mathcal{L}(\mathbf{w}_0) \ge \frac{\alpha}{2}U_0$ holds at $\mathbf{w}_0$. By the initialization condition, $C_0 = \beta\sqrt{\frac{2\mathcal{L}(\mathbf{w}_0)}{\alpha}} \ge \beta\sqrt{U_0}$, thus statement (1) holds. The initial parameter $\mathbf{w}_0$ is the center of the ball, so $\mathbf{w}_0 \in \mathcal{B}_{R/2}(\mathbf{w}_0)$ holds by definition.

Inductive Step ($t \to t+1$): Assume that statements (1) and (2) hold for all steps up to t. The No-Clipping guarantee at step t determines the update rule.

Under interpolation, $\nabla\ell_i$ vanishes at every point of the solution set $\mathcal{S}$: for any $\bar{\mathbf{w}}_t \in \mathcal{S}$, we have $\nabla\ell_i(\bar{\mathbf{w}}_t) = 0$ for all $i \in [n]$. Applying the β-smoothness of ℓ_i between $\mathbf{w}_t$ and $\bar{\mathbf{w}}_t$ (Assumption 3.1), we obtain:

$$\max_{i\in[n]} \|\nabla\ell_i(\mathbf{w}_t)\| = \max_i \|\nabla\ell_i(\mathbf{w}_t) - \nabla\ell_i(\bar{\mathbf{w}}_t)\| \le \beta\|\mathbf{w}_t - \bar{\mathbf{w}}_t\| = \beta\sqrt{U_t} \le C_t.$$

This establishes the No-Clipping guarantee at step t.

Since no clipping occurs, the parameter update becomes $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\nabla\mathcal{L}(\mathbf{w}_t) - \frac{\eta}{n}\zeta^{(t)}$. We now prove statement (1) for $t+1$. Let $\mathbf{w}'_{t+1} = \mathbf{w}_t - \eta\nabla\mathcal{L}(\mathbf{w}_t)$. Using Lemma 3.2, Jensen's inequality (which yields $\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2 \le \mathbb{E}_{i\in[n]}[\|\nabla\ell_i(\mathbf{w}_t)\|^2] \le 2\beta\mathcal{L}(\mathbf{w}_t)$), the Aiming condition, and Proposition F.1, we have:

$$\begin{aligned}\|\mathbf{w}'_{t+1} - \bar{\mathbf{w}}_t\|^2 &= \|\mathbf{w}_t - \bar{\mathbf{w}}_t\|^2 - 2\eta\langle\nabla\mathcal{L}(\mathbf{w}_t), \mathbf{w}_t - \bar{\mathbf{w}}_t\rangle + \eta^2\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2 \\ &\le U_t - 2\eta\theta\mathcal{L}(\mathbf{w}_t) + 2\eta^2\beta\mathcal{L}(\mathbf{w}_t) \\ &= U_t - 2\eta(\theta - \beta\eta)\mathcal{L}(\mathbf{w}_t) \\ &\le U_t - \eta(\theta - \beta\eta)\alpha U_t \\ &= (1-\gamma)U_t. \end{aligned} \tag{F.1}$$

Let $v_t = \mathbf{w}'_{t+1} - \bar{\mathbf{w}}_t$. Expanding the squared parameter distance yields:

$$\begin{aligned}U_{t+1} &\le \left\|v_t - \frac{\eta}{n}\zeta^{(t)}\right\|^2 \\ &\le (1-\gamma)U_t + \left\|\frac{\eta}{n}\zeta^{(t)}\right\|^2 - \frac{2\eta}{n}\langle v_t, \zeta^{(t)}\rangle.\end{aligned}$$

The Gaussian Cross Term: Conditioned on the state at step t, the vector v_t is fixed. Since the injected noise $\zeta^{(t)}$ is drawn from an isotropic Gaussian $\mathcal{N}(0, C_t^2\sigma^2\boldsymbol{I}_p)$, its projection onto the fixed direction v_t forms a one-dimensional zero-mean Gaussian random variable with variance $\nu^2 = \|v_t\|^2C_t^2\sigma^2$. Setting the two-sided Gaussian tail bound $\mathbb{P}(|X| > \tau) \le 2\exp(-\tau^2/(2\nu^2))$ to a failure probability of $\delta_c/(2T)$ yields the threshold $\tau = \nu\sqrt{2\ln(4T/\delta_c)}$. Conditioned on the global success event $\mathcal{E}$, this bound holds for step t, giving:

$$-\frac{2\eta}{n}\langle v_t, \zeta^{(t)}\rangle \le \left|\frac{2\eta}{n}\langle v_t, \zeta^{(t)}\rangle\right| \le \frac{2\eta}{n}\|v_t\|C_t\sigma\sqrt{2\ln(4T/\delta_c)}. \tag{F.2}$$

Substituting this bound, the noise condition $\|\zeta^{(t)}\| \le KC_t$, and the inductive bound $\|v_t\| \le \sqrt{1-\gamma}\sqrt{U_t} \le \sqrt{1-\gamma}\frac{C_t}{\beta}$, we obtain:

$$\begin{aligned}U_{t+1} &\le (1-\gamma)U_t + \left(\frac{\eta}{n}KC_t\right)^2 + \frac{2\eta}{n}\|v_t\|C_t\sigma\sqrt{2\ln(4T/\delta_c)} \\ &\le (1-\gamma)\frac{C_t^2}{\beta^2} + \left(\frac{\eta}{n}KC_t\right)^2 + \frac{2\eta}{n}\left(\sqrt{1-\gamma}\frac{C_t}{\beta}\right)C_t\sigma\sqrt{2\ln(4T/\delta_c)} \\ &= \left((1-\gamma) + \left(\frac{\eta K\beta}{n}\right)^2 + \frac{2\eta\sigma\beta}{n}\sqrt{1-\gamma}\sqrt{2\ln(4T/\delta_c)}\right)\frac{C_t^2}{\beta^2} \\ &\le q^2\frac{C_t^2}{\beta^2} = \frac{C_{t+1}^2}{\beta^2}.\end{aligned}$$

The final inequality holds by Lemma F.2, which completes the proof of statement (1) for $t+1$.

We now prove statement (2) for $t+1$. For each prior step $j \in [0, t]$, the step size is bounded as:

$$\begin{aligned}\|\mathbf{w}_{j+1} - \mathbf{w}_j\| &= \left\|\eta\nabla\mathcal{L}(\mathbf{w}_j) + \frac{\eta}{n}\zeta^{(j)}\right\| \\ &\le \eta C_j + \frac{\eta K C_j}{n} \\ &= \eta\left(1 + \frac{K}{n}\right)C_j.\end{aligned}$$

Summing these step sizes from $j = 0$ to t:

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}_0\| &\leq \sum_{j=0}^{t} \|\mathbf{w}_{j+1} - \mathbf{w}_j\| \\ &\leq \eta \left(1 + \frac{K}{n}\right) C_0 \sum_{j=0}^{t} q^j \\ &< \eta \left(1 + \frac{K}{n}\right) \frac{C_0}{1-q} \\ &\leq \frac{R}{2}.\end{aligned}$$

Thus, $\mathbf{w}_{t+1} \in \mathcal{B}_{R/2}(\mathbf{w}_0)$, proving statement (2) for $t+1$. Both statements hold for $t+1$, completing the induction.

Finally, we remark that the bound ($\sqrt{U_t} \leq \frac{C_t}{\beta}$) established for all $t \in [0, T]$ serves as the mathematical prerequisite for the No-Clipping property. As demonstrated in the inductive step, this bound, combined with β-smoothness under interpolation, guarantees $\max_{i\in[n]} \|\nabla \ell_i(\mathbf{w}_t)\| \leq C_t$ for all $t \in [0, T-1]$. Thus, the No-Clipping and Non-Escape properties hold simultaneously, concluding the proof of Proposition F.3. □

F.3 Proof of Theorem 4.4

Proof. Let $\bar{\mathbf{w}}_T$ be the projection of $\mathbf{w}_T$ onto the optimal set $\mathcal{S}$. In the interpolation regime, we have $\mathcal{L}(\bar{\mathbf{w}}_T) = 0$ and $\nabla\mathcal{L}(\bar{\mathbf{w}}_T) = \mathbf{0}$. By applying the β-smoothness of $\mathcal{L}$ at $\bar{\mathbf{w}}_T$, we have:

$$\mathcal{L}(\mathbf{w}_T) \leq \mathcal{L}(\bar{\mathbf{w}}_T) + \langle \nabla\mathcal{L}(\bar{\mathbf{w}}_T), \mathbf{w}_T - \bar{\mathbf{w}}_T\rangle + \frac{\beta}{2}\|\mathbf{w}_T - \bar{\mathbf{w}}_T\|^2 = \frac{\beta}{2} U_T.$$

Applying $U_T \leq \left(\frac{C_T}{\beta}\right)^2$ with $C_T = C_0 q^T$, $C_0 = \beta\sqrt{\frac{2\mathcal{L}(\mathbf{w}_0)}{\alpha}}$, and $q^2 = 1 - \frac{\gamma}{2}$:

$$\mathcal{L}(\mathbf{w}_T) \leq \frac{\beta}{2}\left(\frac{C_T}{\beta}\right)^2 = \frac{C_0^2}{2\beta} q^{2T} \leq \frac{\beta}{\alpha}\mathcal{L}(\mathbf{w}_0)\left(1 - \frac{\gamma}{2}\right)^T.$$

Using the inequality $(1-x)^k \leq e^{-xk}$ for $x \in (0, 1)$, we bound the geometric decay term by $(1 - \frac{\gamma}{2})^T \leq \exp(-\frac{\gamma}{2}T)$. Substituting the hyperparameter values specified in Theorem 4.4 (yielding $\eta = \tilde{\Theta}\left(\frac{n^2\epsilon^2}{pT}\right)$ and $T = \tilde{\Theta}(n^2\epsilon^2)$) establishes $\gamma = \Theta(\eta) = \tilde{\Theta}\left(\frac{n^2\epsilon^2}{pT}\right)$. Substituting this into the decay term, the final empirical risk is bounded by:

$$\mathcal{L}(\mathbf{w}_T) \leq \frac{\beta}{\alpha} \exp\left(-\tilde{\Theta}\left(\frac{n^2\epsilon^2}{p}\right)\right) \mathcal{L}(\mathbf{w}_0).$$

This completes the proof. □

G Proof of Theorem 5.1

We apply the Gaussian Differential Privacy (GDP) framework to bound the privacy cost of ND-DP-SGD under Poisson subsampling, building upon the analysis in Appendix E.

Proof. While Dong et al. (2022) provides an exact privacy limit for the Poisson subsampled Gaussian mechanism, it involves an intractable integral. We derive a closed-form bound using the χ^2-divergence approach from Bu et al. (2020). Specifically, for a base Gaussian mechanism with single-step parameter $\mu_0 = \frac{2}{\sigma}$ and sampling rate q_s, the total GDP parameter after T steps satisfies:

$$\mu_{total} \leq q_s\sqrt{T\left(e^{4/\sigma^2} - 1\right)}$$

Assuming $\sigma \geq 2$, we have $\frac{4}{\sigma^2} \leq 1$. Using the inequality $e^x - 1 \leq (e-1)x$ for $x \in [0,1]$, we obtain:

$$\mu_{total} \leq q_s \sqrt{T\left((e-1)\frac{4}{\sigma^2}\right)} = \frac{2q_s\sqrt{(e-1)T}}{\sigma}$$

The (ϵ, δ)-DP guarantee holds when this upper bound satisfies the privacy measure conversion constraint (Inequality E.1):

$$\frac{2q_s\sqrt{(e-1)T}}{\sigma} \leq \sqrt{c^2+2\epsilon} - c$$

Solving for σ by multiplying the right-hand side by $(\sqrt{c^2+2\epsilon}+c)$ yields:

$$\sigma \geq \frac{2q_s\sqrt{(e-1)T}(\sqrt{c^2+2\epsilon}+c)}{2\epsilon} = \frac{q_s\sqrt{(e-1)T}(\sqrt{c^2+2\epsilon}+c)}{\epsilon}$$

We now apply the same algebraic relaxations as in Appendix E. For any $c > 0$ and $\epsilon > 0$, we have $\sqrt{c^2+2\epsilon} < c + \frac{\epsilon}{c}$. Under the condition $\epsilon \leq c^2$, it follows that $\frac{\epsilon}{c} \leq c$. Substituting these bounds yields:

$$\frac{q_s\sqrt{(e-1)T}(\sqrt{c^2+2\epsilon}+c)}{\epsilon} < \frac{q_s\sqrt{(e-1)T}\left(2c+\frac{\epsilon}{c}\right)}{\epsilon} \leq \frac{3cq_s\sqrt{(e-1)T}}{\epsilon}$$

For $\delta \leq 0.5$, we have $c = \Phi^{-1}(1-\delta) \geq 0$. By the Chernoff bound for the standard normal distribution, the tail probability is bounded by $\mathbb{P}(Z > c) \leq e^{-c^2/2}$. By the definition of the CDF, $\mathbb{P}(Z > c) = 1 - \Phi(c) = \delta$. Thus, we have $\delta \leq e^{-c^2/2}$, which yields the upper bound $c \leq \sqrt{2\ln(1/\delta)}$. Substituting this into the inequality, we obtain:

$$\frac{3cq_s\sqrt{(e-1)T}}{\epsilon} \leq \frac{3\sqrt{2\ln(1/\delta)}q_s\sqrt{(e-1)T}}{\epsilon} = \frac{q_s\sqrt{18(e-1)T\ln(1/\delta)}}{\epsilon}$$

Thus, the final requirement for σ is:

$$\sigma \geq \frac{q_s\sqrt{18(e-1)T\ln(1/\delta)}}{\epsilon}$$

□

H Proof of Theorem 5.3

In this section, we extend our framework to the stochastic setting, namely Noise-Decay Differentially Private Stochastic Gradient Descent (ND-DP-SGD). We demonstrate that the SGD sampling error and the DP noise under Poisson subsampling can both be bounded and integrated into our geometric envelope while maintaining exponential convergence.

H.1 Bounding the SGD Sampling Error

Lemma H.1 (SGD Sampling Error Bound under Poisson Subsampling)**.** *Assuming a Poisson subsampling rate $q_s = B_s/n$ and that $\max_i \|\nabla \ell_i(\mathbf{w}_t)\| \leq C_t$, let $\xi_s^{(t)} = \frac{1}{B_s}\sum_{i=1}^n I_i \nabla \ell_i(\mathbf{w}_t) - \nabla \mathcal{L}(\mathbf{w}_t)$ denote the SGD sampling error. Then with probability at least $1 - \delta_s/T$, this sampling error satisfies:*

$$\|\xi_s^{(t)}\| \leq C_t\left(\frac{\ln(2T/\delta_s)}{3B_s} + \sqrt{\frac{\ln^2(2T/\delta_s)}{9B_s^2} + \frac{2\ln(2T/\delta_s)}{B_s}}\right) \triangleq C_t K_{SGD}$$

Proof. Using independent Bernoulli indicators $I_i \sim \text{Bernoulli}(q_s)$, the SGD sampling error can be written as a sum of n independent terms:

$$\xi_s^{(t)} = \frac{1}{B_s}\sum_{i=1}^n I_i \nabla \ell_i(\mathbf{w}_t) - \frac{1}{n}\sum_{i=1}^n \nabla \ell_i(\mathbf{w}_t) = \frac{1}{B_s}\sum_{i=1}^n (I_i - q_s)\nabla \ell_i(\mathbf{w}_t)$$

Define the independent zero-mean random vectors $Y_i = (I_i - q_s)\nabla \ell_i(\mathbf{w}_t)$. The norm of each term is bounded by:

$$\|Y_i\| = |I_i - q_s| \|\nabla \ell_i(\mathbf{w}_t)\| \leq \max(q_s, 1 - q_s) C_t \leq C_t$$

The total variance parameter is bounded by evaluating the Bernoulli variance:

$$\sigma^2 = \sum_{i=1}^{n} \mathbb{E}[\|Y_i\|^2] = \sum_{i=1}^{n} q_s(1 - q_s) \|\nabla \ell_i(\mathbf{w}_t)\|^2 \leq n q_s C_t^2 = B_s C_t^2$$

For independent zero-mean random vectors Y_i satisfying $\|Y_i\| \leq M$ and $\sum_{i=1}^{n} \mathbb{E}[\|Y_i\|^2] \leq \sigma^2$, Pinelis' Bernstein Inequality (Pinelis, 1994, Theorem 3.4) states that for $S_n = \sum_{i=1}^{n} Y_i$:

$$\mathbb{P}\left(\|S_n\| \geq \lambda\right) \leq 2 \exp\left(-\frac{\lambda^2}{2(\sigma^2 + M\lambda/3)}\right)$$

Applying this theorem with our upper bounds $M = C_t$ and $\sigma^2 = B_s C_t^2$, and setting $S_n = B_s \xi_s^{(t)}$ with $\lambda = B_s x$, we obtain:

$$\mathbb{P}\left(\|\xi_s^{(t)}\| \geq x\right) \leq 2 \exp\left(-\frac{(B_s x)^2}{2(B_s C_t^2 + C_t B_s x/3)}\right) = 2 \exp\left(-\frac{B_s x^2}{2C_t^2 + \frac{2}{3} C_t x}\right)$$

Equating the right-hand side to δ_s / T yields the quadratic equation:

$$B_s x^2 - \frac{2}{3} C_t \ln(2T/\delta_s) x - 2C_t^2 \ln(2T/\delta_s) = 0$$

Solving for the positive root of x completes the proof. □

H.2 Joint Induction Proof for ND-DP-SGD

In this subsection, we formalize the convergence guarantee of ND-DP-SGD using a joint induction framework, following the analysis applied in the full-batch setting.

Proposition H.2. *Let K_{SGD} denote the sampling error factor defined in Lemma H.1, and let $K_{DP} = \sqrt{p} + \sqrt{2\ln(2T/\delta_c)}$ denote the Gaussian noise bound factor. Under the same conditions as Theorem 5.3, with probability at least $1 - \delta_c - \delta_s$, the following properties hold simultaneously for all $t \in [0, T]$:*

1. $\sqrt{U_t} \leq \frac{C_t}{\beta}$ *(where $U_t \triangleq \|\mathbf{w}_t - \bar{\mathbf{w}}_t\|^2$).*
2. $\|\mathbf{w}_t - \mathbf{w}_0\| \leq \frac{R}{2}$.
3. $\max_{i \in [n]} \|\nabla \ell_i(\mathbf{w}_t)\| \leq C_t$.

Proof. Let $\mathcal{E}_1$ be the event where the DP noise bound and the SGD sampling error bound hold for all $t \in [0, T-1]$ (with joint probability $\geq 1 - \frac{\delta_c}{2} - \delta_s$). Additionally, let $\mathcal{E}_2$ be the event that the Gaussian cross-term projection bound holds across all T steps. As we will show, the projection of the isotropic noise $\zeta_s^{(t)}$ onto the fixed direction at step t admits a failure probability of $\frac{\delta_c}{2T}$ per step. By the union bound, $\mathbb{P}(\mathcal{E}_2) \geq 1 - \frac{\delta_c}{2}$. We define the global success event as $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2$, which guarantees $\mathbb{P}(\mathcal{E}) \geq 1 - \delta_c - \delta_s$.

Conditioned on $\mathcal{E}$, the standard Gaussian vector satisfies $\|\zeta_s^{(t)}\| \leq K_{\text{DP}}$, the SGD sampling error satisfies $\|\xi_s^{(t)}\| \leq C_t K_{\text{SGD}}$, and the cross-term projection bounds hold for all t. We let the local region radius be bounded by $R = \frac{2\eta C_0}{1-q}\left(1 + K_{\text{SGD}} + \frac{\sigma K_{\text{DP}}}{B_s}\right)$.

Base Case ($t = 0$): Proposition F.1 guarantees that the Quadratic Growth condition $\mathcal{L}(\mathbf{w}_0) \geq \frac{\alpha}{2} U_0$ holds at $\mathbf{w}_0$. By the definition of C_0, we have $C_0 = \beta\sqrt{\frac{2\mathcal{L}(\mathbf{w}_0)}{\alpha}} \geq \beta\sqrt{U_0}$, thus $\sqrt{U_0} \leq \frac{C_0}{\beta}$. Under interpolation, $\nabla \ell_i(\bar{\mathbf{w}}_0) = 0$ for all $i \in [n]$. Applying β-smoothness yields $\max_i \|\nabla \ell_i(\mathbf{w}_0)\| \leq \beta \|\mathbf{w}_0 - \bar{\mathbf{w}}_0\| = \beta\sqrt{U_0} \leq C_0$. Finally, the initial distance is $\|\mathbf{w}_0 - \mathbf{w}_0\| = 0 \leq \frac{R}{2}$. Thus, the base case holds.

Inductive Step ($t \to t+1$)**:** Assume that the three properties hold for all iterations $k \le t$.

Property 1: By the inductive hypothesis $\max_i \|\nabla \ell_i(\mathbf{w}_t)\| \le C_t$, no clipping occurs at step t. Thus, the mini-batch gradient sum can be decomposed into the exact full gradient and the SGD sampling error: $\frac{1}{B_s}\sum_{i\in\mathcal{B}_t} \tilde{g}_i^{(t)} = \frac{1}{B_s}\sum_{i\in\mathcal{B}_t} \nabla \ell_i(\mathbf{w}_t) = \nabla\mathcal{L}(\mathbf{w}_t) + \xi_s^{(t)}$. Substituting this into the update rule, expanding the squared distance, and applying the non-expansive property of the projection onto $\mathcal{S}$:

$$\begin{aligned} U_{t+1} &= \|\mathbf{w}_{t+1} - \bar{\mathbf{w}}_{t+1}\|^2 \le \|\mathbf{w}_{t+1} - \bar{\mathbf{w}}_t\|^2 \\ &= \left\| \mathbf{w}_t - \bar{\mathbf{w}}_t - \eta\left(\nabla\mathcal{L}(\mathbf{w}_t) + \xi_s^{(t)}\right) - \frac{\eta C_t \sigma}{B_s}\zeta_s^{(t)} \right\|^2. \end{aligned}$$

Let $v_t = \mathbf{w}_t - \bar{\mathbf{w}}_t - \eta\nabla\mathcal{L}(\mathbf{w}_t)$ and define the pre-noise update vector $v_t' = v_t - \eta\xi_s^{(t)}$. Since the standard Gaussian noise $\zeta_s^{(t)}$ is drawn independently after the mini-batch sampling, v_t' is fixed with respect to $\zeta_s^{(t)}$. Expanding the square yields:

$$U_{t+1} \le \|v_t'\|^2 + \left\| \frac{\eta C_t \sigma}{B_s}\zeta_s^{(t)} \right\|^2 - \frac{2\eta C_t\sigma}{B_s}\langle v_t', \zeta_s^{(t)}\rangle.$$

From the gradient contraction in Inequality F.1, we have $\|v_t\| \le \sqrt{1-\gamma}\sqrt{U_t}$. Applying the triangle inequality and the SGD sampling error bound $\|\xi_s^{(t)}\| \le C_t K_{\text{SGD}}$, the norm of the combined vector is bounded by:

$$\|v_t'\| \le \|v_t\| + \eta\|\xi_s^{(t)}\| \le \sqrt{1-\gamma}\sqrt{U_t} + \eta C_t K_{\text{SGD}} \le \left(\frac{\sqrt{1-\gamma}}{\beta} + \eta K_{\text{SGD}}\right) C_t,$$

where the final step utilizes the inductive hypothesis $\sqrt{U_t} \le \frac{C_t}{\beta}$.

Following the same argument as for Inequality F.2, applying the two-sided Gaussian tail bound with a failure probability of $\delta_c/(2T)$ yields the high-probability threshold. Conditioned on $\mathcal{E}$, the projection of the standard Gaussian vector $\zeta_s^{(t)}$ onto the fixed direction v_t' is bounded by:

$$-\frac{2\eta C_t\sigma}{B_s}\langle v_t', \zeta_s^{(t)}\rangle \le \frac{2\eta C_t\sigma}{B_s}\|v_t'\|\sqrt{2\ln(4T/\delta_c)}.$$

Substituting the bounds for $\|v_t'\|$, bounding the squared standard Gaussian noise by $\|\zeta_s^{(t)}\|^2 \le K_{\text{DP}}^2$ (following Lemma B.1), and collecting terms yields:

$$\begin{aligned} U_{t+1} &\le \left[\left(\sqrt{1-\gamma} + \eta\beta K_{\text{SGD}}\right)^2 + \left(\frac{\eta\beta\sigma K_{\text{DP}}}{B_s}\right)^2 \right. \\ &\qquad \left. + \frac{2\eta\beta\sigma}{B_s}\left(\sqrt{1-\gamma} + \eta\beta K_{\text{SGD}}\right)\sqrt{2\ln(4T/\delta_c)} \right] \frac{C_t^2}{\beta^2} \qquad \text{(H.1)} \\ &\le q^2 \frac{C_t^2}{\beta^2} = \frac{C_{t+1}^2}{\beta^2}, \end{aligned}$$

The final inequality holds under the required decay rate condition (which is verified in the Final Proof Assembly below), establishing the bound for step $t+1$.

Property 2: Using the update rule and our noise bounds:

$$\|\mathbf{w}_{k+1} - \mathbf{w}_k\| \le \eta\|\nabla\mathcal{L}(\mathbf{w}_k)\| + \eta\|\xi_s^{(k)}\| + \frac{\eta C_k \sigma K_{\text{DP}}}{B_s} \le \eta\left(1 + K_{\text{SGD}} + \frac{\sigma K_{\text{DP}}}{B_s}\right) C_k$$

Summing these step sizes from $k = 0$ to t:

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}_0\| &\leq \sum_{k=0}^{t} \|\mathbf{w}_{k+1} - \mathbf{w}_k\| \\ &\leq \sum_{k=0}^{t} \eta \left(1 + K_{\text{SGD}} + \frac{\sigma K_{\text{DP}}}{B_s}\right) C_k \\ &= \eta C_0 \left(1 + K_{\text{SGD}} + \frac{\sigma K_{\text{DP}}}{B_s}\right) \sum_{k=0}^{t} q^k \\ &\leq \eta \left(1 + K_{\text{SGD}} + \frac{\sigma K_{\text{DP}}}{B_s}\right) \frac{C_0}{1-q} = \frac{R}{2}\end{aligned}$$

Thus, the training trajectory remains within the local region.

Property 3: Applying the β-smoothness of ℓ_i between $\mathbf{w}_{t+1}$ and $\bar{\mathbf{w}}_{t+1}$ (Assumption 3.1):

$$\|\nabla \ell_i(\mathbf{w}_{t+1})\| = \|\nabla \ell_i(\mathbf{w}_{t+1}) - \nabla \ell_i(\bar{\mathbf{w}}_{t+1})\| \leq \beta \|\mathbf{w}_{t+1} - \bar{\mathbf{w}}_{t+1}\| \leq \beta \sqrt{U_{t+1}} \leq C_{t+1}$$

This guarantees that for every sample $i \in [n]$, the gradient norm is bounded by C_{t+1}. This concludes the joint induction step. □

H.3 Final Proof Assembly

We now derive the hyperparameter conditions required to bound the expansion in Inequality H.1 by $q^2 = 1 - \frac{\gamma}{2}$.

We decouple the requirement by allocating the budget $1 - \frac{\gamma}{2}$ across the three additive components of the expansion. For the first term, expanding the square yields:

$$\left(\sqrt{1-\gamma} + \eta\beta K_{\text{SGD}}\right)^2 = (1-\gamma) + 2\sqrt{1-\gamma}\eta\beta K_{\text{SGD}} + (\eta\beta K_{\text{SGD}})^2.$$

We set this expression to be bounded by $1 - \frac{3\gamma}{4}$, which implies that the remainder $2\sqrt{1-\gamma}\eta\beta K_{\text{SGD}} + (\eta\beta K_{\text{SGD}})^2$ must be bounded by $\frac{\gamma}{4}$. Since the first-order term $2\sqrt{1-\gamma}\eta\beta K_{\text{SGD}}$ dominates this remainder, this condition is satisfied when $\eta\beta K_{\text{SGD}} = \mathcal{O}(\gamma) = \mathcal{O}(\eta(\theta - \beta\eta)\alpha)$. Canceling η on both sides yields:

$$\beta K_{\text{SGD}} = \beta \left(\frac{\ln(2T/\delta_s)}{3B_s} + \sqrt{\frac{\ln^2(2T/\delta_s)}{9B_s^2} + \frac{2\ln(2T/\delta_s)}{B_s}} \right) \leq \mathcal{O}\big((\theta - \beta\eta)\alpha\big).$$

Since $K_{\text{SGD}} = \mathcal{O}\left(\sqrt{\frac{\ln(T)}{B_s}}\right)$, this implies the batch size lower bound $B_s = \Omega(\ln T)$.

The remaining budget $\frac{\gamma}{4}$ is distributed equally between the second and third terms in Inequality H.1. Substituting $\sigma = \frac{q_s\sqrt{18(e-1)T\ln(1/\delta)}}{\epsilon}$ and $q_s/B_s = 1/n$ yields:

$$\frac{\sigma}{B_s} = \frac{\sqrt{18(e-1)T\ln(1/\delta)}}{n\epsilon}.$$

For the second term, we require:

$$\left(\frac{\eta\beta\sigma K_{\text{DP}}}{B_s}\right)^2 \leq \frac{\gamma}{8} = \frac{\eta(\theta - \beta\eta)\alpha}{8}.$$

Substituting $\frac{\sigma K_{\text{DP}}}{B_s} = \tilde{\mathcal{O}}\left(\frac{\sqrt{Tp}}{n\epsilon}\right)$, we obtain the learning rate bound:

$$\eta \leq \tilde{\mathcal{O}}\left(\frac{n^2\epsilon^2}{pT}\right).$$

For the third term, we require:

$$\frac{2\eta\beta\sigma}{B_s}(\sqrt{1-\gamma}+\eta\beta K_{\text{SGD}})\sqrt{2\ln(4T/\delta_c)} \leq \frac{\gamma}{8} = \frac{\eta(\theta-\beta\eta)\alpha}{8}.$$

Applying the first term's constraint $\sqrt{1-\gamma}+\eta\beta K_{\text{SGD}} \leq \sqrt{1-\frac{3\gamma}{4}} < 1$ and substituting $\frac{\sigma}{B_s}\sqrt{2\ln(4T/\delta_c)} = \tilde{\mathcal{O}}\left(\frac{\sqrt{T}}{n\epsilon}\right)$ yields the iteration bound:

$$T \leq \tilde{\mathcal{O}}\left(n^2\epsilon^2\right).$$

To satisfy these bounds simultaneously, we set the learning rate as $\eta = \min\left\{\frac{\theta}{2\beta}, \frac{\theta\alpha B_s^2}{16\beta^2\sigma^2 K_{\text{DP}}^2}\right\}$ and the total iterations as $T = \tilde{\Theta}(n^2\epsilon^2)$, which matches the hyperparameter configuration defined in Theorem 5.3.

Using the distance bound $U_T \leq \left(\frac{C_T}{\beta}\right)^2$ at step T with $C_T = C_0 q^T$, $C_0 = \beta\sqrt{\frac{2\mathcal{L}(\mathbf{w}_0)}{\alpha}}$, and $q^2 = 1-\frac{\gamma}{2}$:

$$\mathcal{L}(\mathbf{w}_T) \leq \frac{\beta}{2}U_T \leq \frac{\beta}{2}\left(\frac{C_T}{\beta}\right)^2 = \frac{C_0^2}{2\beta}q^{2T} \leq \frac{\beta}{\alpha}\mathcal{L}(\mathbf{w}_0)\left(1-\frac{\gamma}{2}\right)^T.$$

Using the inequality $(1-x)^k \leq e^{-xk}$ for $x \in (0,1)$, we bound the geometric decay term by $(1-\frac{\gamma}{2})^T \leq \exp(-\frac{\gamma}{2}T)$. Substituting the hyperparameter values specified in Theorem 5.3 (yielding $\eta = \tilde{\Theta}\left(\frac{n^2\epsilon^2}{pT}\right)$ and $T = \tilde{\Theta}(n^2\epsilon^2)$) establishes $\gamma = \Theta(\eta) = \tilde{\Theta}\left(\frac{n^2\epsilon^2}{pT}\right)$. Substituting this into the decay term, the final empirical risk is bounded by:

$$\mathcal{L}(\mathbf{w}_T) \leq \frac{\beta}{\alpha}\exp\left(-\tilde{\Theta}\left(\frac{n^2\epsilon^2}{p}\right)\right)\mathcal{L}(\mathbf{w}_0).$$

H.4 Requirements on the Subsampling Rate

Finally, we verify the assumption $\sigma \geq 2$ required in Theorem 5.1. Substituting the required noise scale $\sigma = \frac{q_s\sqrt{18(e-1)T\ln(1/\delta)}}{\epsilon}$ yields:

$$q_s \geq \frac{2\epsilon}{\sqrt{18(e-1)T\ln(1/\delta)}}$$

Using $B_s = nq_s$ and substituting $T = \tilde{\Theta}(n^2\epsilon^2)$, we obtain:

$$B_s \geq \tilde{\Omega}\left(n\cdot\frac{\epsilon}{\sqrt{n^2\epsilon^2}}\right) = \tilde{\Omega}(1).$$

Since this requirement evaluates to $\tilde{\Omega}(1)$, the batch size lower bound is determined by the stronger condition $B_s = \Omega(\ln T)$ derived above for the sampling error.

I Experiments

We experimentally evaluate the ND-DP-GD framework on the MNIST and CIFAR-2 datasets. All experiments were executed on a single NVIDIA H100 GPU. Executing the complete suite of comparative experiments, which encompasses 5 random seeds, various privacy budgets, and all hyperparameter variations, takes approximately 5 minutes for the MNIST task and 29 minutes for the CIFAR-2 task.

I.1 Experiments on MNIST

Experimental Setup.

Dataset and Model. We use a binary subset of the MNIST dataset (digits 0 and 1) (LeCun et al., 1998), which is distributed under the Creative Commons Attribution-Share Alike 3.0 license. This

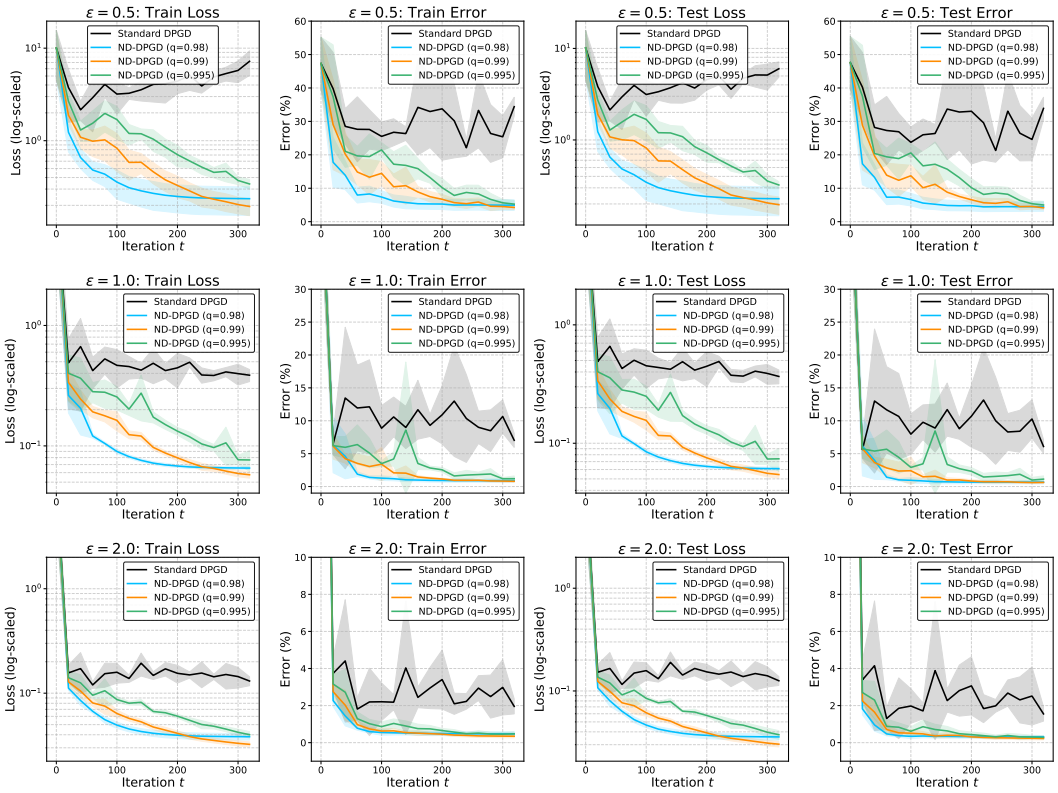


Figure 1: Experimental results of DP-GD and ND-DP-GD on MNIST. From left to right: training loss curves (log-scaled), empirical risk, test loss curves (log-scaled), test error. Rows: various privacy noise bounds σ. ND-DP-GD is tested under different settings of decay rates (q). Each curve is obtained by averaging over 5 independent runs. The shaded regions represent one standard deviation. The results are consistent with the theoretical prediction: under interpolation, synchronized decay converges faster and achieves a lower empirical risk, while fixed noise restricts DP-GD to a high loss floor.

subset provides $n = 12,665$ training samples and $2,115$ test samples. We train a fully connected neural network with two hidden layers, of widths 128 and 64, respectively. The weights are initialized using Kaiming initialization, and the biases are initialized to zero. Each hidden layer is activated using the Softplus activation function with its scale parameter set to 5. We use the Mean Squared Error (MSE) loss for the optimization. The total number of parameters in our model is $p = 108,866$.

Algorithm Settings. For standard DP-GD, we set the clipping threshold to $C = 50$ and learning rate to $\eta = 0.01$. For ND-DP-GD, we use the same learning rate and initialize the clipping threshold to $C_0 = 50$. We test three decay rates: $q \in \{0.98, 0.99, 0.995\}$ for ND-DP-GD. For both algorithms, we run $T = 320$ epochs using full-batch gradients. We evaluate the performance under privacy parameters $\epsilon \in \{0.5, 1.0, 2.0\}$ and $\delta = 1/n$. The noise scaling factor σ is chosen according to Theorem 4.4 so that differential privacy is guaranteed by Theorem 3.7. In each setting, we make 5 independent runs with different seeds and plot the average and standard deviation.

Results. The experimental results are reported in Figure 1. We observe that, after the initial phase of training, the empirical loss of standard DP-GD (black curves) oscillates around a relatively high floor. This observation aligns with our theoretical understanding: in the interpolation regime, static noise accumulation eventually oversizes the diminishing gradients, causing the empirical risk to stay high. In contrast, the trajectories with synchronized decay (curves of other colors) avoid this accumulation, converging faster and reaching a much lower empirical risk (shown on the logarithmic y-axes). This is consistent with our theory that coupling the decay mechanism with the automatic gradients reduction phenomenon yields a much lower empirical risk.

Additional experiments and discussions using a convolutional neural network on the CIFAR dataset are provided in Appendix I.2 below.

I.2 Experiments on CIFAR-2

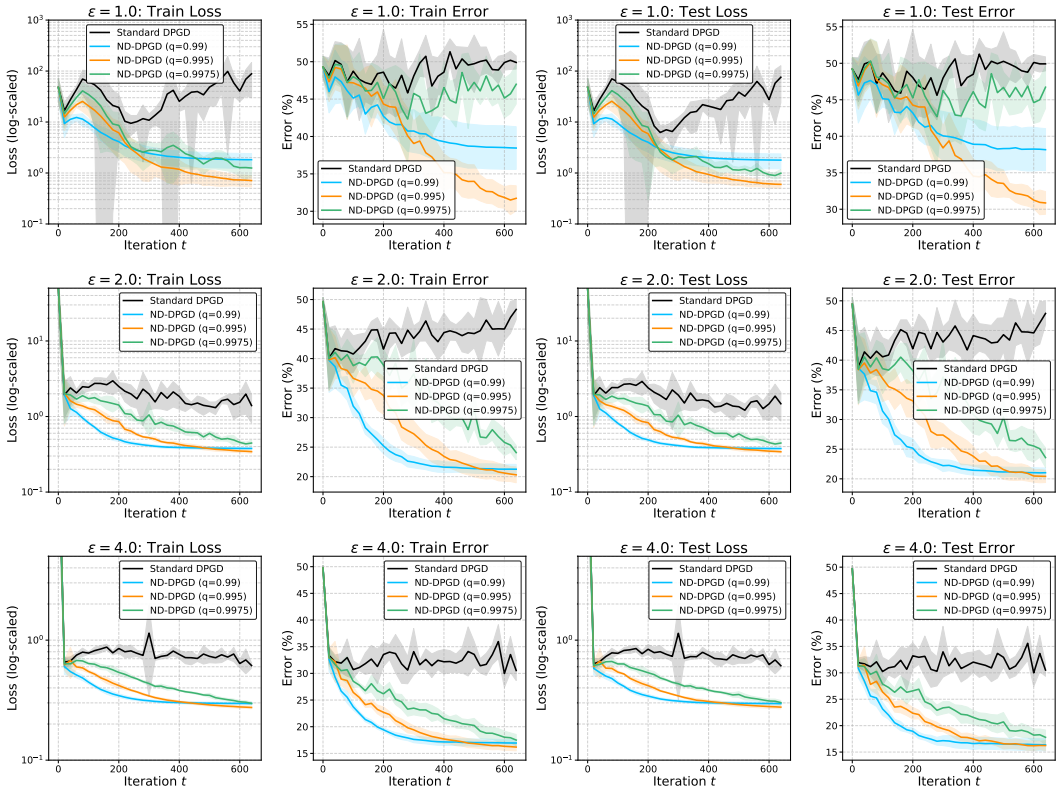


Figure 2: Experimental results of DP-GD and ND-DP-GD on CIFAR-2. Columns left to right: training loss curves (log-scaled), empirical risk, test loss curves (log-scaled), test error. Rows: various privacy budgets (ϵ). ND-DP-GD is tested under different settings of decay rates (q). Each curve is obtained by averaging over 5 independent runs. The shaded regions represent one standard deviation.

Experimental Setup.

Dataset and Model. We create a binary classification task on CIFAR-2 by extracting the "airplane" and "automobile" classes from the CIFAR-10 dataset (Krizhevsky et al., 2009), which is made available under the MIT license. This provides $n = 10,000$ training samples and $2,000$ test samples.

We train a CNN specifically structured as follows. All layer weights are initialized using Kaiming initialization, and biases are initialized to zero.

- **Convolutional Block 1:** Conv2d (3 input channels, 16 output channels, kernel size 3×3, padding 1), followed by ReLU activation and MaxPool2d (kernel size 2×2, stride 2).
- **Convolutional Block 2:** Conv2d (16 input channels, 32 output channels, kernel size 3×3, padding 1), followed by ReLU activation and MaxPool2d (kernel size 2×2, stride 2).
- **Fully Connected Block:** Flatten operation, Linear layer (from $32 \times 8 \times 8 = 2048$ dimensions to 128 dimensions), ReLU activation, and a final Linear output layer (from 128 dimensions to 2 dimensions).

We use the Mean Squared Error (MSE) loss for the optimization. The total number of parameters is $p = 267,618$.

Algorithm Settings. We extend the training process to $T = 640$ epochs and reduce the learning rate to $\eta = 0.005$ for both algorithms. For standard DP-GD, we set the clipping threshold to $C = 150$. For ND-DP-GD, we set the initial clipping threshold to $C_0 = 150$ and test three decay rates: $q \in \{0.99, 0.995, 0.9975\}$. We evaluate the performance under privacy parameters $\epsilon \in \{1.0, 2.0, 4.0\}$ and $\delta = 1/n$. σ is chosen according to Theorem 4.4 so that differential privacy is guaranteed by Theorem 3.7. In each setting, we also have 5 independent runs with different random seeds and report the average and standard deviation.

Results. The training and testing dynamics are presented in Figure 2. Similar to the MNIST evaluations, the empirical loss of standard DP-GD (black line) stays high after an initial decrease across all privacy budgets. In contrast, ND-DP-GD converges fast and reaches much lower empirical risk, when q is appropriately set. This confirms the effectiveness of our proposed noise-decay mechanism in ND-DP-GD even in the over-parameterized non-convex neural network setting.